Machine Learning A.A. 2023-2024
Progetto Finale

Churn Analysis: un approccio data-driven per il miglioramento delle performance aziendali

Autore: Luca De Simone

Abstract

Nel settore delle telecomunicazioni, la previsione del churn rappresenta una sfida strategica fondamentale per le aziende, in quanto consente di implementare azioni mirate per ridurre la perdita di clienti e ottimizzare l'allocazione delle risorse aziendali. In questo studio, viene applicata un'analisi avanzata di machine learning per identificare i fattori che influenzano la disdetta del contratto tramite l'utilizzo di diversi algoritmi quali Regressione Logistica, Support Vector Machine a base radiale, Reti Neurali, Alberi classificativi e Random Forest.

L'integrazione di questi modelli all'interno di un framework costi-profitti consente di massimizzare l'efficacia delle strategie di retention, migliorando la fidelizzazione e ottimizzando la spesa del budget aziendale. Inoltre, lo sviluppo di sistemi di classificazione delle motivazioni di churn e la segmentazione di nuovi clienti rispetto agli storici possono offrire insight data-driven ancora più approfonditi, supportando decisioni volte a potenziare le strategie di marketing e customer care.

L'analisi permette di dimostrare come un approccio predittivo basato sull'analisi dei dati possa fornire un vantaggio competitivo significativo, permettendo alle compagnie di anticipare le tendenze della clientela e di adottare delle misure proattive finalizzate all'affermazione dell'azienda all'interno del mercato e al suo sviluppo a lungo termine.

Indice

1. Introduzione

Il customer churn è un problema sempre più rilevante nel mondo aziendale e, nell'era digitale, rappresenta un elemento cruciale per il successo e la sostenibilità delle imprese. Questo fenomeno, particolarmente diffuso nei settori basati su contratti e abbonamenti, si verifica quando un cliente decide di rescindere la propria sottoscrizione passando ad un competitor o rinunciando del tutto al servizio.

L'abbandono clientelare impatta significativamente sulle economie aziendali comportando una perdita diretta delle entrate, una riduzione della quota di mercato e un aumento dei costi per l'acquisizione di nuovi consumatori in sostituzione di quelli persi. Studi recenti dimostrano che un'impresa, per attrarre un nuovo cliente, può arrivare a spendere una quota fino a 5-7 volte superiore rispetto a quella necessaria alla fidelizzazione di uno già esistente. Per tale ragione, sempre più aziende pongono la customer retention al centro delle loro strategie in quanto hanno compreso il ruolo centrale che i clienti hanno nella crescita di quest'ultime. Con l'avvento dei social, i consumatori sono diventati essi stessi dei brand ambassadors e la loro capacità di diffondere opinioni rapidamente rappresenta un enorme potenziale economico in termini di fatturato, posizionamento sui mercati e di brand identity da parte delle imprese.

Per comprendere il comportamento dei clienti al fine di ridurre al minimo la loro propensione al churning, e quantificare con esattezza quanto questa influirà sul bilancio aziendale, molte imprese si affidano all'analisi dei big data. Tuttavia**,** rilevare il churning è una sfida complessa per diverse ragioni. Innanzitutto, l'identificazione dei fattori scatenanti e la valutazione del loro impatto è molto difficoltoso, le cause possono essere molteplici, spesso client-specific, e legate un'insoddisfazione verso il servizio, ad un'assistenza inadeguata o alla concorrenza, capace di offrire condizioni più vantaggiose o esperienze più soddisfacenti. In secondo luogo, il churn è un evento raro, il che significa che la quota di clienti che abbandonano in un determinato periodo è generalmente minoritaria rispetto alla totalità. Questo crea uno squilibrio nei dati, rendendo difficile la costruzione di modelli predittivi accurati, poiché tenderanno sempre a favorire la classe dominante, ovvero coloro che non hanno intenzione di disdire il servizio. Infine, la grande quantità di dati può rappresentare sia un'opportunità che un limite. L'abbondanza informativa spesso include dati non rilevanti e sovrapposizioni, che se non trattate in modo ottimale rendono complicata l'estrazione di insight significativi.

L'obiettivo dell'analisi è applicare metodologie di machine learning a dati reali di un'azienda californiana di telecomunicazioni, per sviluppare modelli automatici capaci d'identificare e prevedere i potenziali abbandoni della clientela, così da ridurne il tasso e massimizzarne i profitti. Per raggiungere tale scopo, saranno implementate diverse tecniche tra cui: **Random Forest (RF), Support Vector Machine (SVM), Regressione Logistica (LR) e Neural Networks (NN)**.

2. Struttura dei dati e selezione delle osservazioni per l'analisi del churn

Le compagnie telefoniche hanno a disposizione una grande quantità di dati sui loro clienti, tra cui informazioni demografiche, utilizzo dei servizi, durata della sottoscrizione, reclami e storico dei pagamenti. Analizzando queste informazioni, è possibile individuare pattern e segnali utili a identificare possibili churners al fine di adottare strategie proattive di fidelizzazione per ridurne il tasso.

Per tale ragione un'azienda di telecomunicazioni californiana, che per motivi di privacy rimarrà ignota, ha messo a disposizione all'interno di una data challenge disponibile sulla piattaforma di Maven Analytics, una porzione dei propri dati relativa al secondo trimestre del 2022.

Il dataset contiene informazioni su 7.043 clienti e ogni record rappresenta un consumatore dal punto di vista delle sue generalità, durata della sottoscrizione, servizi attivati e il suo stato corrente, ovvero se si tratta di nuovo cliente, di uno già esistente oppurre di uno che ha appena disdetto il proprio abbonamento. Nello specifico si hanno a disposizione 38 variabili, nove relative a informazioni personali come genere, stato civile e residenza, mentre le restanti riguardanti il contratto pattuito, servizi a cui il cliente ha accesso, modalità di pagamento e storico delle spese.

Ai fini dell'analisi sono state considerate solamente le osservazioni inerenti ai churners e agli stayed, escludendo così i nuovi clienti. Questo perchè, avendo appena sottoscritto il proprio abbonamento, non hanno potuto esercitare una scelta arbitraria sul restare all'interno della compagnia o abbandonarla. Per tale ragione, il numero totale di individui sulla quale è stata esecitata l'analisi sarà pari a 6589.

Analizzando la frequenza delle osservazioni all'interno delle categorie churned e stayed si è osservato come, nel secondo trimestre del 2022, la quota di soggetti che hanno revocato il proprio abbonamento è circa pari al 28,36 %, minoritaria rispetto alla percentuale di clienti che ha rinnovato, ma comunque rilevante.

3. Pre-Processing ed Analisi Esplorative

3.1 Missing Values e Ricodifica delle Variabili

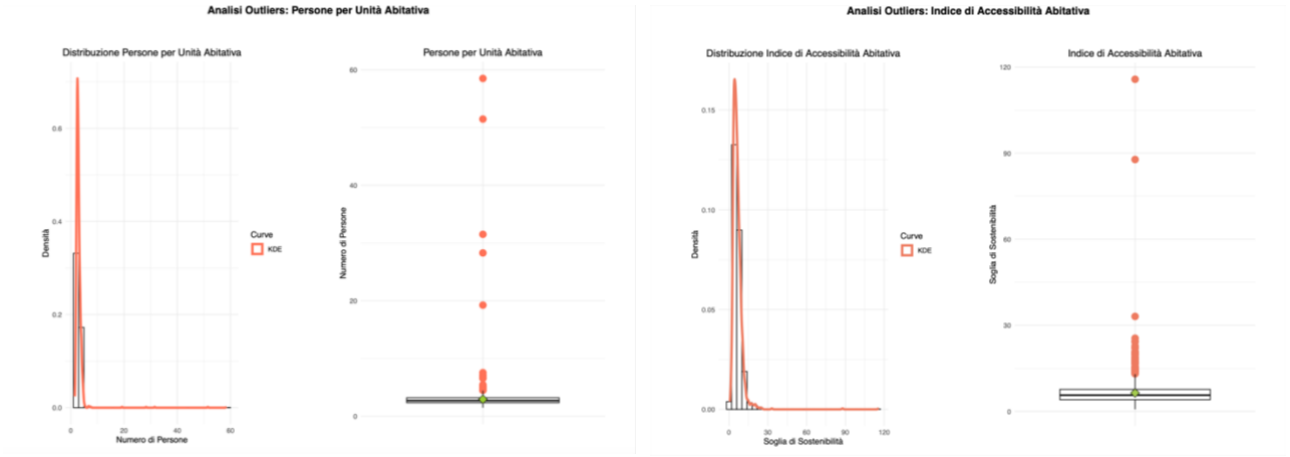
I dati in esame non contengono valori mancanti in senso stretto, ma presentano delle istanze missing a causa della natura delle variabili. Nei contratti di abbonamento, come quelli offerti dalle compagnie di telecomunicazioni, sono spesso previste opzioni vincolate che, se accettate, consentono l'accesso a dei servizi aggiuntivi. Gli utenti che scelgono di non includere tali opzioni nel loro contratto avranno, quindi, dei valori mancanti nelle voci relative ai servizi extra, i quali sono stati ricodificati in *"opzione non disponibile"*. Le variabili vincolanti che inducono la problematica appena descritta sono **Abbonamento telefonico** e **Abbonamento Internet,** le quali indicano, rispettivamente, se il cliente è abbonato al servizio fisso telefonico e alla rete internet con la compagnia.

3.2 Metodi di clustering per effettuare optimal grouping

Analizzando le variabili, si è osservato che le features **CAP** e **Città** presentano una perfetta sovrapposizione informativa. Inoltre, si è notato che entrambe evidenziano un numero molto elevato di modalità, rispettivamente pari a 1106 e 1626, che, se non trattato, potrebbe risultare problematico in fase di stima dei modelli. Per tale ragione, si è deciso di escludere dall'analisi la covariata **Città** in quanto collineare e con una granularità informativa minore. Successivamente, mediante il pacchetto R *"zipcodeR",* è stato possibile estrapolare, a partire dal **CAP**, delle informazioni aggiuntive delle singole aree urbane come popolazione residente, estensione territoriale relativa, unità abitatative totali e occupate, reddito residente e residenziale mediano. Quest'ultime sono risultate fondamentali, insieme alla Latitudine e Longitudine, nell'identificare delle macroaree territoriali utili per ridurre il numero delle modalità ed eseguire un'optimal grouping.

Per realizzare il raggruppamento sono stati utilizzati diversi algoritmi di clustering, tra cui **K-Means**, **H-clust** e **DBScan**. In una prima fase, sono state costruite due nuove variabili **Densità Popolativa** e **Tasso di Occupazione Residenziale**, rispettivamente numero di residenti per supperficie territoriale e percenutale di case occupate sulla totalità delle unità abitative. Queste, insieme ai valori di

Latitudine, **Longitudine**, **Reddito Residenziale** e **Residente** mediano, hanno permesso di costituire un dataset per la clusterizzazione dei CAP sulla base delle caratteristiche sociodemografiche delle singole aree territoriali. Da un'analisi effettuata è emerso come alcune features presentassero diversi valori mancanti e anomali, identificati tramite le variabili di controllo **Persone per Unità Abitativa** e **Indice di Accessibilità Abitativa**. La prima, ha permesso di verificare se vi fossero delle incongruenze o anomalie all'interno dei dati come una sottostima della popolazione o del numero di unità abitative occupate. La seconda, consiste in una metrica in grado di stimare la sostenibilità finanziaria dell'acquisto di un'abitazione in relazione al reddito annuale del nucleo familiare. Secondo la letteratura, una casa risulta sostenibile se il suo prezzo è inferiore a tre volte il reddito annuale personale. Per la variabile, trattandosi di una stima basata su quantità di sintesi e avendo a disposizione solamente il valore delle abitazioni, e non la spesa sostenuta per l'acquisto, è stata considerata una soglia di sostenibilità pari a 20, oltre la quale un'osservazione è da considerarsi anomala.

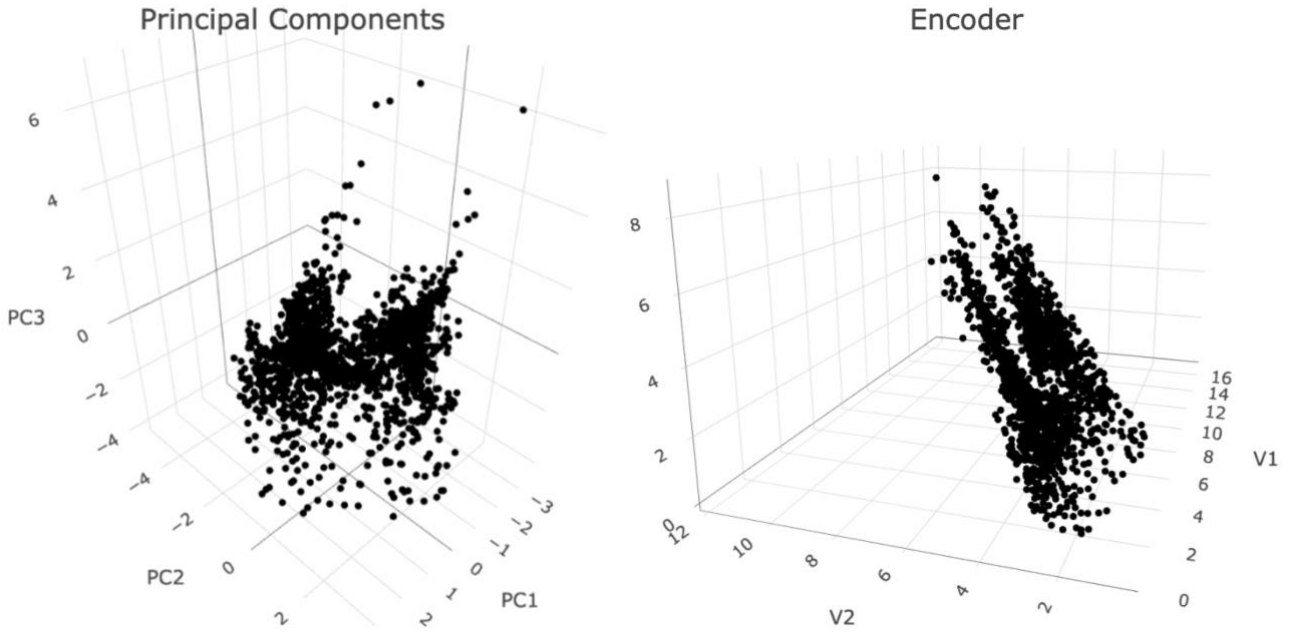


I grafici riportati in figura hanno permesso di evidenziare gli outliers rilevati dalle due features e, non sapendo quale caratteristica ne fosse la responsabile, si è optato per porre degli *"NA"* in corrispondenza delle variabili coinvolte. Successivamente, sono state eseguite venti **Random Forest**, ciascuna con un'inizializzazione diversa, che hanno consentito, attraverso la mediana delle singole stime, l'imputazione dei dati mancanti e la risuluzione delle anomalie identificate.

3.2.1 Riduzione della dimensionalità e definizione delle metriche di distanza

Per applicare i metodi di clustering sopracitati, trattandosi di algoritmi basati sul concetto di distanza, è stato necessario il calcolo di quest'ultime. Siccome le caratteristiche selezionate per il raggruppamento sono tutte quantità numeriche con ordini di grandezze e unità di misure differenti, per renderle confrontabili, è stata eseguita una standardizzazione. Inoltre, considerando un numero di variabili elevato che non permette la visualizzazione dello spazio congiunto, e nota la sofferenza dei metodi alla maledizione della dimensionalità, sono state utilizzate diverse accortezze. La prima è stata quella di calcolare, oltre alla classica **distanza euclidea**, anche una metrica adattiva che permettesse di contemplare la struttura e la densità locale dei dati, ovvero la **distanza di Mahalanobis**. La seconda è stata quella di eseguire in parallelo un clustering che prendesse in input, anziché le features originali, una loro versione ottenuta mediante l'applicazioni di due metodi di riduzione della dimensionalità quali **Componenti Principali** e **Reti Neurali**. Quest'ultime, se configurate come **AutoEncoder**, permettono, tramite un'architettura a farfalla, di ricostruire gli input codificandoli in una rappresentazione latente di dimensione ridotta che viene poi rispansa per ricomporre le features iniziali. Quindi, considerando solamente la prima parte di codifica delle covariate, l'**Encoder**, è possibile eseguire una data reduction non lineare da confrontare a quella eseguita dalle Principal Components. La riduzione della dimensionalità presenta, però, alcuni effetti collaterali, tra cui il seguente. Ridurre lo spazio di input da p dimensioni a uno di output con k dimensioni, con $k < p$,

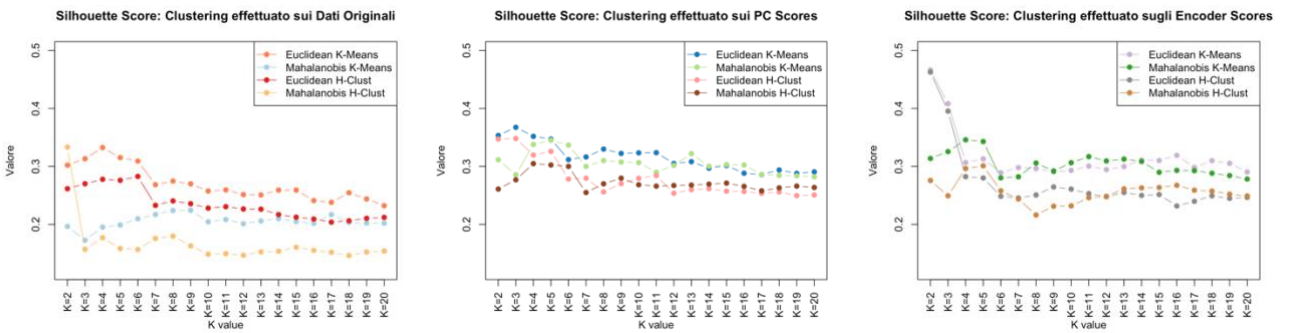
potrebbe portare a una forte concentrazione dei punti, alterando la struttura originaria dello spazio. In particolare, potrebbe accadere che punti inizialmente distanti nello spazio di input risultino molto vicini in output e viceversa. Se la riduzione dimensionale è eccessiva, si rischia di ottenere un insieme di punti fortemente addensato, perdendo gran parte delle relazioni spaziali iniziali. Osservando gli scatter plot delle nuove covariate ottenute, si nota esattamente questo fenomeno, nonostante si è passati da uno spazio di input a sei dimensioni ad uno di output di tre.



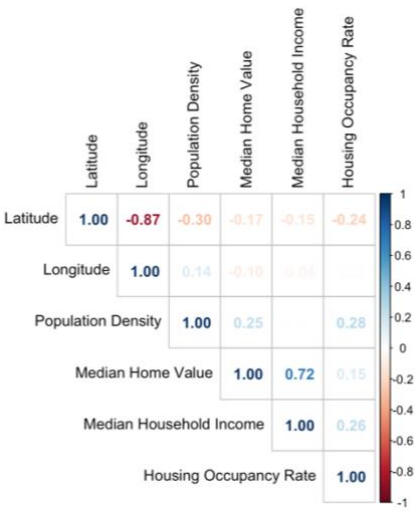
La scelta di ridurre lo spazio delle covariare a tre dimensioni è stata motivata principalmente per una ragione di visualizzazione e poi per preservare una quantità significativa di informazioni senza introdurre forti distorsioni e bias. Infatti, nel caso delle componenti principali ha permesso di spiegare l'82,38% della variabilità totale come riportato in tabella.

Principal Components	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1,4769	1,3177	1,0126	0,8796	0,4636	0,2616
Proportion of Variance	0,3635	0,2894	0,1709	0,1290	0,0358	0,0114
Cumulative Proportion	0,3635	0,6529	0,8238	0,9528	0,9886	1,0000

Questo ha reso impossibile l'applicazione di metodi di clustering basati su densità locali, come il **DBScan**, in quanto l'eccessiva concentrazione delle osservazioni ne ha uniformato la distribuzione impedendo al metodo di identificare correttamente i clusters. Tuttavia, è stato possibile applicare i metodi basati sulle distanze i quali sono risultati concordi nell'identificare la presenza di tre cluster, come evidenziato nei grafici riportati in figura.



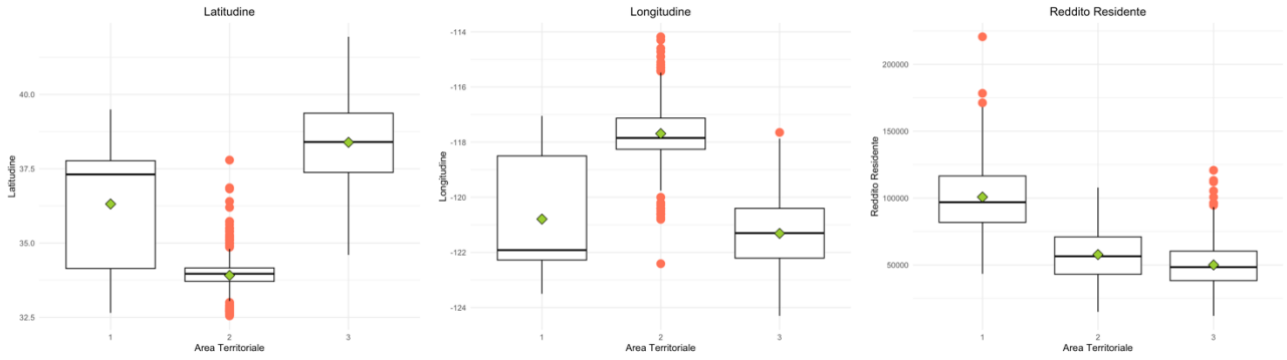
3.2.2 Definizione della metrica di valutazione ed interpretazione dei clusters



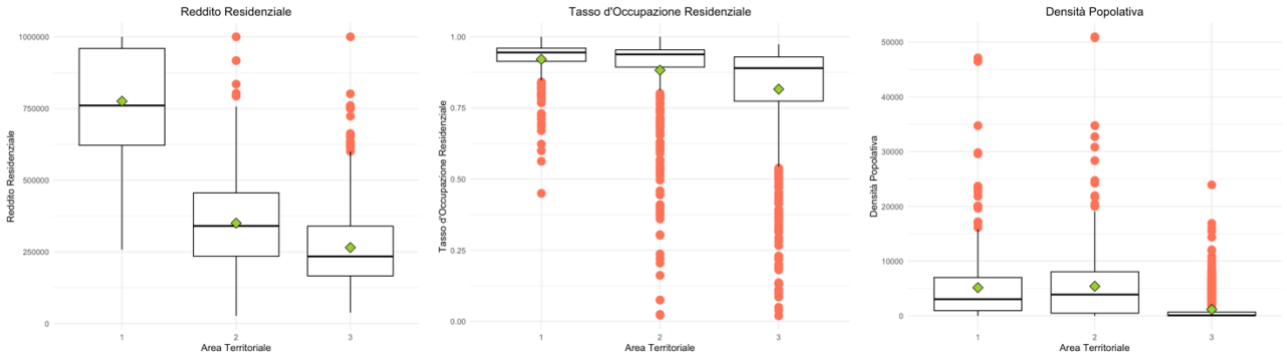
Trattandosi di un problema non supervisionato, la media dei **Silhouette Scores** è stata utilizzata come misura di validazione interna dei risultati. Questa metrica risulta particolarmente utile per confrontare diverse soluzioni di clustering in quanto adimensionale, il suo valore è compreso tra **-1** e **1** indipendentemente dalla scala o dalle caratteristiche numeriche dei dati. Valori negativi indicano che le osservazioni sono state mal classificate, valori vicini allo zero che i punti si trovano al confine tra più gruppi e, valori prossimi ad uno denotano una buona assegnazione ai clusters.
Analizzando i risultati non si evidenziano differenze significative tra i clustering ottenuti utilizzando le distanze euclidee e quelle di Mahalanobis. Questo perché la distanza di Mahalanobis è una versione normalizzata della distanza euclidea in cui si tiene conto della matrice di covarianza delle variabili. Quando le features presentano correlazioni deboli o trascurabili, la matrice di covarianza si avvicina ad una matrice diagonale, rendendo la distanza di Mahalanobis prossima a quella euclidea. Questo comportamento è stato confermato dall'analisi del correlogramma riportato.

In seguito, è stato eseguito un **ensemble** delle diverse metodologie di clustering e di pre-processing mediante ***majority vote***. Questa strategia ha permesso di ottenere soluzioni più robuste e generalizzabili, in quanto la fusione di più tecniche consente di ridurre l'instabilità e la sensibilità dei risultati. Inoltre, l'adozione di un ensemble consente di mitigare i rischi derivanti da possibili errori o incertezze introdotti da un singolo algoritmo, migliorando così l'affidabilità complessiva delle previsioni e rendendo la segmentazione robusta alle fluttuazioni dei dati di input.

In conclusione, si è proceduto ad interpretare i tre cluster così costituiti. I grafici riportati in seguito hanno evidenziato come le coordinate geografiche di una determinata area territoriale siano risultate fortemente significative nella segmentazione dei CAP.



Inoltre, sulla base delle variabili socioeconomiche è stato possibile la profilazione delle tre macro-zone. La terza rappresenta l'insieme dei territori rurali, caratterizzati da un reddito residente e residenziale nettamente al di sotto delle altre due aree, mentre la prima racchiude le zone residenziali più abbienti. Infine, il secondo gruppo rappresenta le aree abitate dal ceto medio contraddistinte da un alto tasso occupazionale e da redditi consistenti.



3.3 Analisi della Near Zero Variance e della Collinearità

Conclusa la fase di optimal grouping, che ha portato l'esclusione delle variabili **CAP** e **Città** in favore di **Macroarea**, si è proceduto ad eseguire un'analisi della **Near Zero Variance**. Questa ha comportato l'eliminazione sia dei **Rimborsi Totali** che delle **Spese Totali Dati Extra**, in quanto pressochè degeneri e dall'apporto informativo nullo, l'89% delle loro osservazioni è risultato avere il medesimo valore.

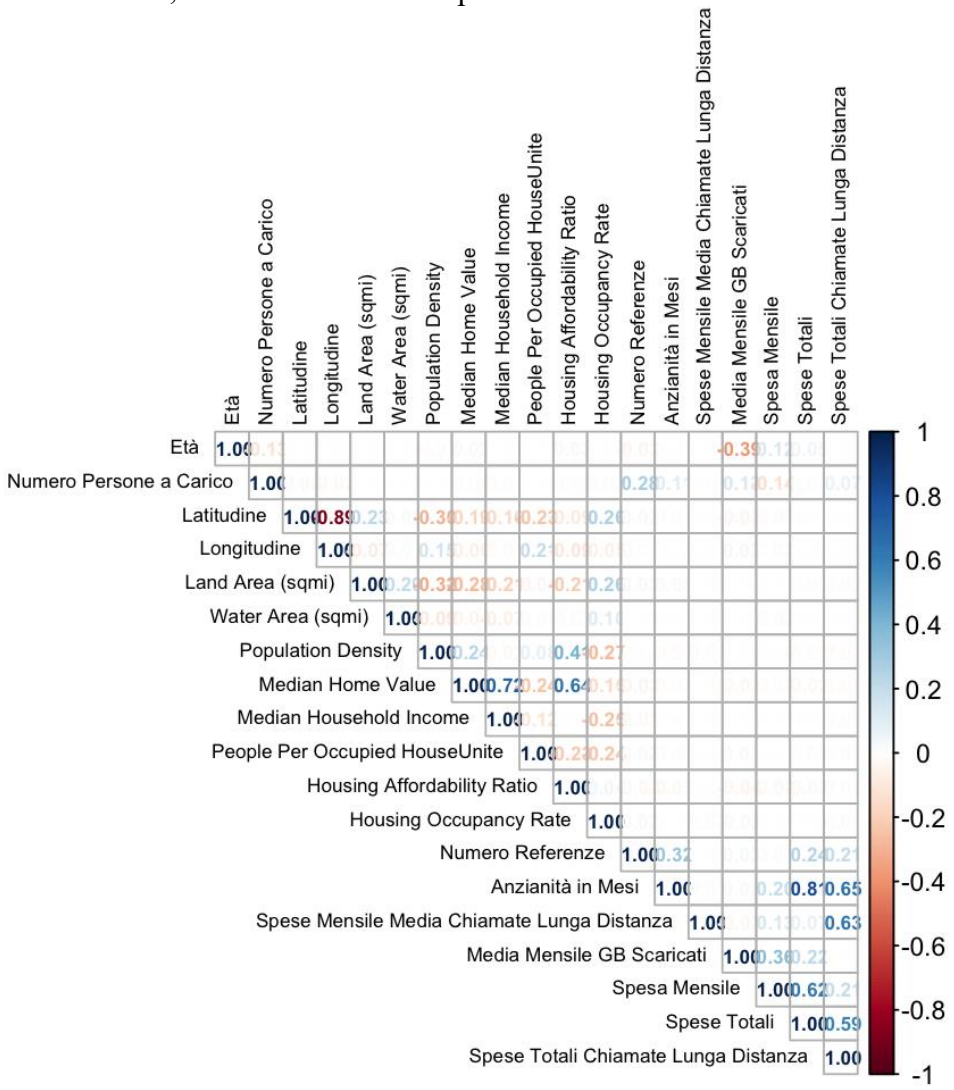
Lo studio delle **collinearità** ha permesso di evidenziare una problematica intriseca ai dati contrattuali. La presenza di opzioni vincolate, nello specifico la feature **Abbonamento Internet**, limita la possibilità dell'utente di sottoscriversi a servizi aggiuntivi connessi a quest'ultima. Da un punto di vista statistico, è possibile affermare che le variabili **Sicurezza online**, **Backup online**, **Piano di protezione dei dispositivi**, **Supporto tecnico premium**, **Streaming TV**, **Streaming film**, **Streaming musica** e **Dati illimitati** siano funzioni della vincolante e quindi fortemente connesse a quest'ultima. La ragione di tale fenomeno è dovuta al fatto che se un utente non ha aderito all'offerta internet, automaticamente non ha avuto la possibilità di accedere ai servizi extra sopracitati, presentando alle relative voci la dicitura *"opzione non disponibile"*. Questo, oltre a rendere le variabili connesse ad **Abbonamento Internet** le rende fortemente associate tra loro. Tale fenomeno è stato possibile evidenziarlo mediante la metrica del **Chi-Quadro** e dell'**indice V di Cramér**, entrambi riportati in tabella.

Variabile	Variabile	Chi Square value	P-Value	Cramér's V
Abbonamento Internet	Sicurezza Online	6.928,18	0	0,725
Abbonamento Internet	Backup Online	6.591,82	0	0,707
Abbonamento Internet	Piano Protezione Disposi	6.590,70	0	0,707
Abbonamento Internet	Supporto Tecnico Premium	6.892,04	0	0,723
Abbonamento Internet	Streaming TV	6.745,87	0	0,716
Abbonamento Internet	Streaming Film	6.732,66	0	0,715
Abbonamento Internet	Streaming Musica	6.624,52	0	0,709
Abbonamento Internet	Dati Illimitati	6.589,57	0	0,707
Backup Online	Sicurezza Online	6.791,29	0	0,718
Sicurezza Online	Supporto Tecnico Premium	7.077,11	0	0,733
Supporto Tecnico Premium	Piano Protezione Disposi	6.935,87	0	0,726
Piano Protezione Disposi	Streaming Film	7.093,85	0	0,734
Streaming Film	Streaming Musica	10.950,99	0	0,912
Streaming Musica	Streaming TV	7.346,78	0	0,747
Streaming TV	Dati Illimitati	6.590,34	0	0,707

Rispettivamente il test del **Chi-Quadro** ha permesso di verificare se esiste una relazione significativa tra due variabili categoriali, mentre la **V di Cramér** ha consentito di misurare la forza della relazione

tra due quest'ultime, dove 0 indica nessuna associazione e 1 perfetta. Per alcuni modelli, tali variabili potrebbero risultare problematiche da gestire, allora si è proceduto, per quello che è stato possibile, ad accorpare i servizi tra loro. Nello specifico, sono state costruite due nuove variabili **Sicurezza** e **Streaming**, le quali, composte rispettivamente da **Sicurezza online, Piano di protezione dei dispositivi** e **Streaming TV**, **Streaming film**, **Streaming musica**, riportano la voce *"abbonato"* anche se un utente avesse sottoscritto ad un solo servizio degli elencati.

In conclusione, si è analizzato il correlogramma, il quale non ha messo in evidenza nessun tipo di anomalia. Le variabili non sembrano avere sovrapposizioni informative, si identificano solamente alcune correlazioni rilevanti, nello specifico quelle tra i valori di Latitudine e Longitudine, Reddito Residente e Residenziale mediano, Anzianità in mesi e Spese Totali.



4. Classificazione: definizione delle metriche e degli obiettivi

L'analisi è stata condotta adottando due approcci distinti, i cui risultati potrebbero non coincidere necessariamente.

Il primo approccio si concentra sulle singole unità e ha l'obiettivo di **massimizzare la customer retention**, ovvero identificare con la massima accuratezza i clienti potenzialmente intenzionati a disdire il proprio abbonamento in modo da poter offrire loro incentivi mirati. Per perseguire tale scopo, i modelli sono stati ottimizzati in base all'**F-Score**, una metrica che rappresenta la media

armonica tra **Sensitivity (Recall)** e **Precision (True Positive Rate)**. Questa scelta è motivata dalla necessità dell'azienda di riconoscere correttamente i **churners**, ottimizzando le risorse destinate alla fidelizzazione dei clienti ed evitando di investire budget su utenti che non hanno alcuna intenzione di recedere dal contratto. In questo contesto la **Sensitivity (Recall)** misura la proporzione di churners che il modello ha identificato correttamente, mentre la **Precision (True Positive Rate)** indica la percentuale di utenti classificati come churners che effettivamente lo sono. Poiché queste due metriche sono inversamente correlate, l'ottimizzazione dell'**F-Score** consente di bilanciare entrambi gli aspetti, migliorando la capacità del modello nell'identificare un cliente come churner solo quando lo è realmente.

Il secondo criterio adottato ha l'obiettivo di **massimizzare i profitti aziendali**, differenziandosi da un'accurata classificazione generale per concentrarsi esclusivamente sull'identificazione dei clienti più profittevoli. In questo contesto, ogni cliente è stato **ponderato** in base ad una variabile creata ad hoc**, Entrate Mensili**, definita come il rapporto tra le **Entrate Totali** generate dal cliente e la sua **Anzianità** contrattuale.
L'idea alla base di questo approccio è quella di **ottimizzare le risorse destinate alla fidelizzazione**, investendole solo sui clienti che hanno un impatto significativo sul fatturato aziendale, evitando sprechi su utenti con un contributo economico marginale.
Per raggiungere tale obiettivo, i modelli sono stati ottimizzati per **massimizzare una funzione di guadagno**, strutturata in modo che se un cliente viene classificato correttamente, il valore delle sue **Entrate Mensili** viene **aggiunto al profitto complessivo**, altrimenti il valore corrispondente viene **sottratto**, penalizzando così gli errori di classificazione.
Questo approccio consente di allineare l'ottimizzazione del modello agli interessi aziendali, privilegiando la **retention dei clienti ad alto valore** e garantendo un utilizzo più efficiente del budget destinato alla riduzione del churn.

Per condurre l'analisi, il set di dati è stato suddiviso in due sottoinsiemi, **traning set** (80%) e **test set** (20%), **preservando la distribuzione originale** della variabile target **Stato Cliente**. Per effettuare lo studio, sono stati adottati cinque modelli di classificazione: **Regressione Logistica, Support Vector Machine con kernel radiale, Albero Decisionale, Random Forest e Rete Neurale**, ognuno dei quali è stato addestrato mediante una **10-fold cross-validation stratificata**. Questo ha permesso di garantire che, in ciascuna suddivisione, la proporzione della classe target rimanesse invariata, in modo da non verificarsi alcun sbilanciamento che avrebbe potuto compromettere la validazione degli iperparametri. In tal senso, l'ottimizzazione degli **iperparametri** è stata effettuata attraverso un approccio **bayesiano basato su processi gaussiani**. L'infill criteria adottato, noto come **Lower Confidence Bound (LCB)** o **GP-Lower Confidence Bound (GP-LCB)**, seleziona, come valore da esplorare, il limite inferiore della banda di incertezza associata alla previsione del processo, in formula:

$$LCB(x) = \mu(x) - \kappa \cdot \sigma(x)$$

dove $\mu(x)$ è la media predetta dal processo gaussiano nel punto x, $\sigma(x)$è la deviazione standard predetta in x e κ è un parametro di esplorazione, che nell'analisi è stato fissato pari a 2, che controlla il trade-off tra exploration ed exploitation in modo da non convergere prematuramente in un minimo locale.

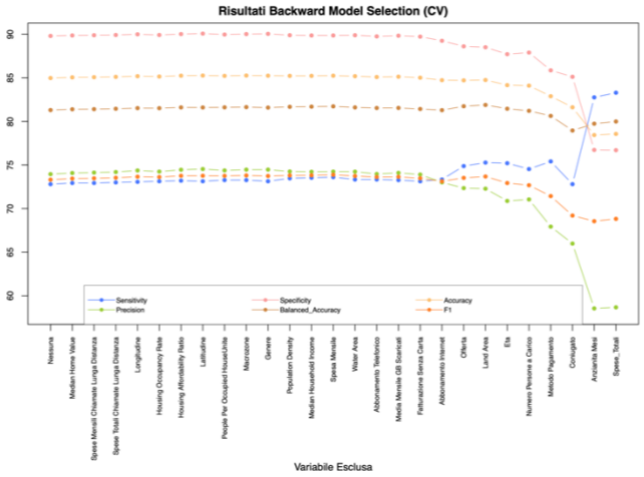
L'uso dell'ottimizzazione bayesiana ha permesso una ricerca efficace dei minimi delle funzioni obiettivo, con una maggior garanzia di convergenza ai punti globali ed una minor suscettibilità a scelte casuali inefficaci, come spesso accade con metodologie quali random search o grid search.

4.1 Modelli imbalanced basati sulle unità

L'addestramento di ciascun predittore è stato effettuato mediante **cross-validation**, suddividendo, ad ogni iterazione, il dataset in **un training set di 4744 osservazioni**, utilizzato per l'apprendimento del modello, e **un test set di 528 osservazioni**, per la valutazione delle performance. Per ragioni computazionali, il numero di combinazioni di iperparametri esplorate durante l'ottimizzazione bayesiana è stato definito **ad hoc** per ciascun classificatore, bilanciando il trade-off tra efficienza e accuratezza della ricerca. Gli iperparametri selezionati e le relative performance di ogni modello sono riportati di seguito.

4.1.1 Modello Logistico

Nel caso del classificatore logistico, l'unico iperparametro da definire è il set di variabili da utilizzare per predire la classe di appartenenza di ciascun cliente. Per selezionare l'insieme di variabili ottimale, è stata adottata una procedura di **backward cross-validation**, finalizzata a individuare il numero minimo di feature in grado di massimizzare l'**F-Score**, bilanciando così le prestazioni classificative con la capacità di generalizzazione del modello. Per evitare problemi di perfetta collinearità, sono stati testati cinque diversi set di variabili, ciascuno dei quali differiva dagli altri esclusivamente per la presenza delle variabili connesse ai servizi aggiuntivi precedentemente individuate, ovvero **Sicurezza, Backup online, Supporto tecnico premium, Streaming e Dati illimitati**. Il set di variabili selezionato come ottimale dalla procedura, insieme alle relative performance, è riportato in seguito.



Modello Logistico Imbalanced			
F-Score	Precision	Sensitivity	Specificity
0,7389262	0,7419137	0,7359626	0,8985699

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1101	383
Stayed	395	3393

4.1.2 Albero di Classificazione

Gli iperparametri tunati mediante ottimizzazione bayesiana sono stati:

- **MinSplit**: numero minimo di osservazioni richieste affinché un nodo possa essere ulteriormente approfondito.
- **CP (Parametro di complessità)**: soglia minima di riduzione dell'errore richiesto per effettuare una suddivisione; se il miglioramento dell'errore è inferiore a questo valore, il nodo non viene diviso.

Durante la validazione incrociata sono state esplorate 1000 combinazioni di iperparametri, individuando come configurazione ottimale quella composta da **MinSplit** = 50 e **CP** = 0,0035.

Albero Classificativo Imbalanced			
F-Score	Precision	Sensitivity	Specificity
0,7392106	0,8246914	0,6697861	0,9435911

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1002	213
Stayed	494	3563

4.1.3 Random Forest

Il classificatore è stato ottimizzato rispetto ai seguenti iperparametri:

- **mtry**: numero di variabili selezionati casualmente a ogni split di un albero della foresta.
- **ntree**: numero totale di alberi generati e combinati per costituire la foresta.

L'ottimizzazione bayesiana, eseguita in cross-validation, ha esplorato 150 combinazioni di iperparametri, identificando la seguente come ottimale **mtry** = 10 e **ntree** = 1433

Random Forest Imbalanced			
F-Score	Precision	Sensitivity	Specificity
0,7478456	0,8508099	0,6671123	0,9536547

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	998	175
Stayed	498	3601

4.1.4 Support Vector Machine con Kernel Radiale

Nel caso delle **SVM con kernel radiale**, l'ottimizzazione bayesiana ha esplorato 300 combinazioni di iperparametri, ognuna composta da:

- **Costo (C)**: parametro di penalità che controlla il trade-off tra bias e varianza, bilanciando la massimizzazione del margine in funzione degli errori di classificazione.
- **Gamma (γ)**: parametro che determina l'influenza dei singoli punti nella definizione della funzione decisionale. Gamma è inversamente proporzionale a **sigma (σ)**, che rappresenta la bandwidth del kernel radiale.

Dalla ricerca ottimizzata, è stata individuata la combinazione di iperparametri più performante costituita da **Costo** = 1073, 826 e **Gamma** = 30,12305

Support Vector Machine Imbalanced			
F-Score	Precision	Sensitivity	Specificity
0,7261324	0,7583697	0,6965241	0,9120763

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1042	332
Stayed	454	3444

4.1.5 Neural Network

Per ragioni computazionali, se è adottata una **rete neurale shallow**, ovvero con un solo **strato nascosto**. Questo perché l'addestramento di una rete deep con validazione incrociata e ottimizzazione bayesiana, basata su processi gaussiani, avrebbe richiesto un costo computazionale troppo oneroso. Per la costruzione della rete neurale è stata utilizzata la libreria R *"nnet"*, che prevede l'ottimizzazione di due iperparametri chiave:

- **Decay**: tasso di regolarizzazione (*weight decay*), che penalizza i pesi eccessivamente grandi al fine di ridurre il rischio di overfitting.
- **Size**: inumero di neuroni presenti nello strato nascosto, influenzando la capacità della rete di apprendere pattern complessi.

In questo caso, l'ottimizzazione bayesiana ha esplorato 100 combinazioni di iperparametri, individuando la configurazione **Decay** = 0.4884 e **Size** = 8 come ottimale

Neural Network Imbalanced			
F-Score	Precision	Sensitivity	Specificity
0,7109727	0,7372577	0,6864973	0,903072

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1027	366
Stayed	469	3410

I risultati ottenuti dai modelli non mostrano valori particolarmente elevati per l'**F-Score**, con un valore medio pari al 73,26%. Tuttavia, analizzando più nel dettaglio, emergono alcune osservazioni interessanti. Se da un lato i modelli presentano una buona **Precision** (ovvero un alto **True Positive Rate**), che indica una forte capacità di classificare correttamente i Churners quando questi vengono identificati, dall'altro la **Sensitivity** (o **Recall**) è relativamente bassa, con una media pari al 69,12%. Questo suggerisce una difficoltà da parte dei modelli nell'identificare tutti i potenziali Churners, ma una volta individuati, tendono a classificarli correttamente con una precisione media del 78,26%. Le ragioni di questo fenomeno potrebbero essere due. In primo luogo, i clienti che hanno intenzione di recedere dal contratto potrebbero non presentare caratteristiche distintive facilmente identificabili, in

quanto la decisione di abbandonare il servizio potrebbe essere fortemente influenzata da fattori **client-specific** piuttosto che da difetti evidenti contrattuali. In secondo luogo, la proporzione di Churners all'interno dei fold di addestramento potrebbe non essere sufficientemente adeguata a permettere ai modelli di apprendere un pattern chiaro e affidabile per la loro identificazione.
Per cercare di migliorare le prestazioni dei classificatori, è stata adottata la seguente strategia. L'addestramento continua ad essere eseguito utilizzando una **10-fold cross-validation stratificata**, in modo da garantire che la distribuzione della classe target rimanga invariata all'interno di ciascun fold. L'ottimizzazione degli iperparametri continua ad avvalersi di un approccio bayesiano basato su processi gaussiani con il criterio di infill **LCB (Lower Confidence Bound)**. Per incrementare la proporzione di **Churners** all'interno dei training fold, senza alterare la distribuzione delle classi nei test fold, il target è stato ribilanciato tramite **oversampling**, impiegando il metodo **ROSE** ad ogni iterazione. Quest'ultima è una tecnica di bilanciamento che genera esempi sintetici per la classe minoritaria, cercando di preservare le caratteristiche strutturali dei dati originali. Piuttosto che duplicare semplicemente le osservazioni della classe minoritaria, ROSE crea nuovi campioni combinando casualmente le istanze delle osservazioni esistenti. Questo processo avviene in modo tale da non alterare la variabilità delle features, con l'obiettivo di produrre nuovi esempi plausibili che rappresentino fedelmente la distribuzione reale dei dati. In questo modo, si evita il problema della sovra-rappresentazione di campioni identici e si favorisce una distribuzione più equilibrata tra le classi, migliorando le capacità dei modelli nella rilevazione dei churners.

4.2 Modelli balanced basati sulle unità

Come avvenuto per i modelli addestrati su dati sbilanciati è stata eseguita un'ottimizzazione bayesiana cross-validata che ha portato all'identificazione dei seguenti iperparametri e risultati.

4.2.1 Modello Logistico

Modello Logistico Balanced			
F-Score	Precision	Sensitivity	Specificity
0,7177465	0,6202532	0,8516043	0,7934322

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1274	780
Stayed	222	2996

4.2.2 Albero di Classificazione

MinSplit = 143 e **CP** = 0,0036.

Albero Classificativo Balanced			
F-Score	Precision	Sensitivity	Specificity
0,6919767	0,5568228	0,9137701	0,7118644

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1367	1088
Stayed	129	2688

4.2.3 Random Forest

mtry = 10 e **ntree** = 1159.

Random Forest Balanced			
F-Score	Precision	Sensitivity	Specificity
0,7068421	0,5828993	0,8977273	0,7454979

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1343	961
Stayed	153	2815

4.2.4 Support Vector Machine con Kernel Radiale

Costo = 1,8365 e **Gamma** = 4,592.

Support Vector Machine Balanced			
F-Score	Precision	Sensitivity	Specificity
0,7124711	0,6270325	0,8248663	0,8056144

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1234	734
Stayed	262	3042

4.2.5 Neural Network

Decay = 0.407 e **size** = 9.

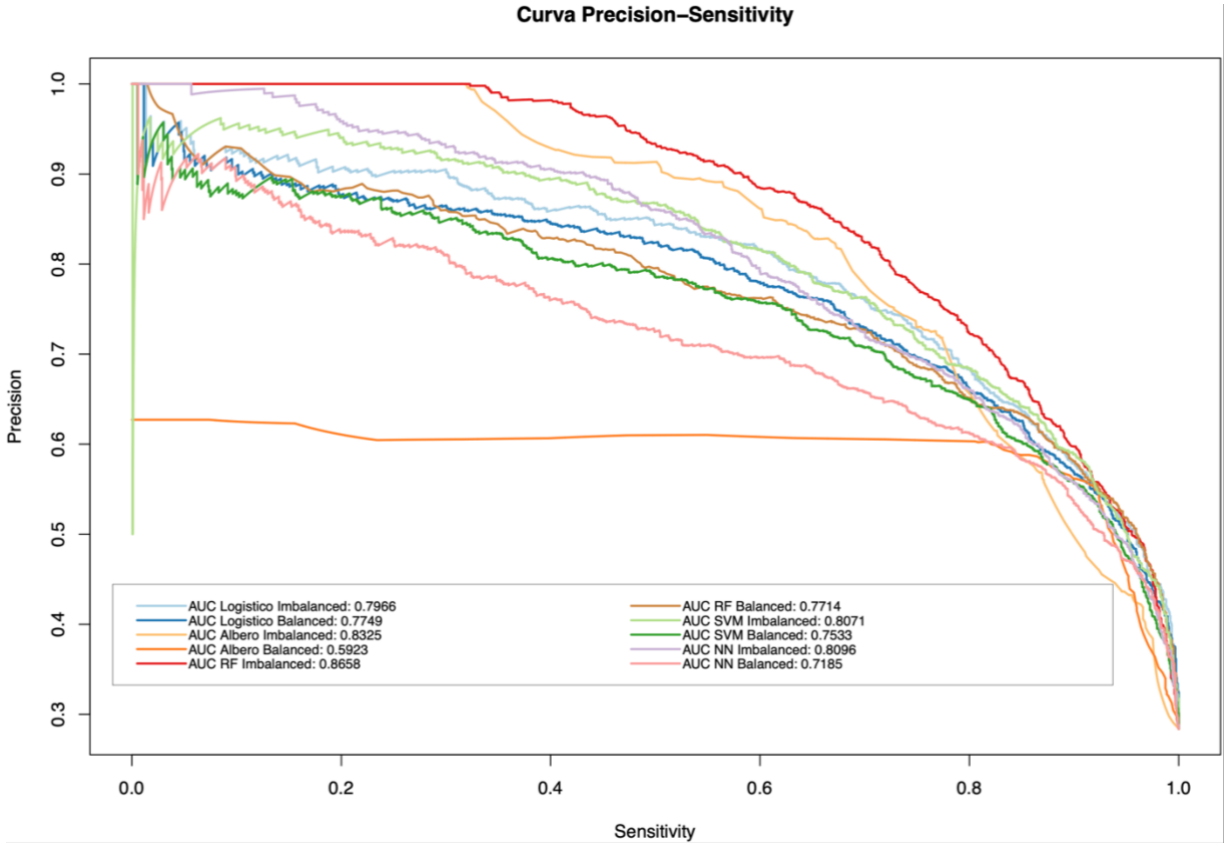
Neural Network Balanced			
F-Score	Precision	Sensitivity	Specificity
0,691044	0,572432	0,8716578	0,7420551

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1304	974
Stayed	192	2802

Si osserva che, dopo il bilanciamento dei dati, i valori di **F-Score** sono rimasti pressoché invariati, si è passato da un 73,26% medio ad un 70,40%. Questo fenomeno è dovuto al fatto che i modelli, sebbene abbiano acquisito una maggiore propensione nel prevedere i churners, hanno contemporaneamente subito una riduzione della loro **precisione**. Rispettivamente, si è passati da una **Sensitivity** media del 69,12% ad una del 87,19% e da una **Precision** di 78,26% al 59,19%.
La causa principale di questo effetto risiede nella relazione inversa tra **Sensitivity** e **Precision**, aumentando la capacità di identificare la classe minoritaria, migliorando la **Recall**, si è verificato un inevitabile calo della **Precision**. In sintesi, il bilanciamento ha contribuito principalmente a migliorare la capacità dei modelli nel rilevare i churners, senza tuttavia incrementare le loro prestazioni complessive.

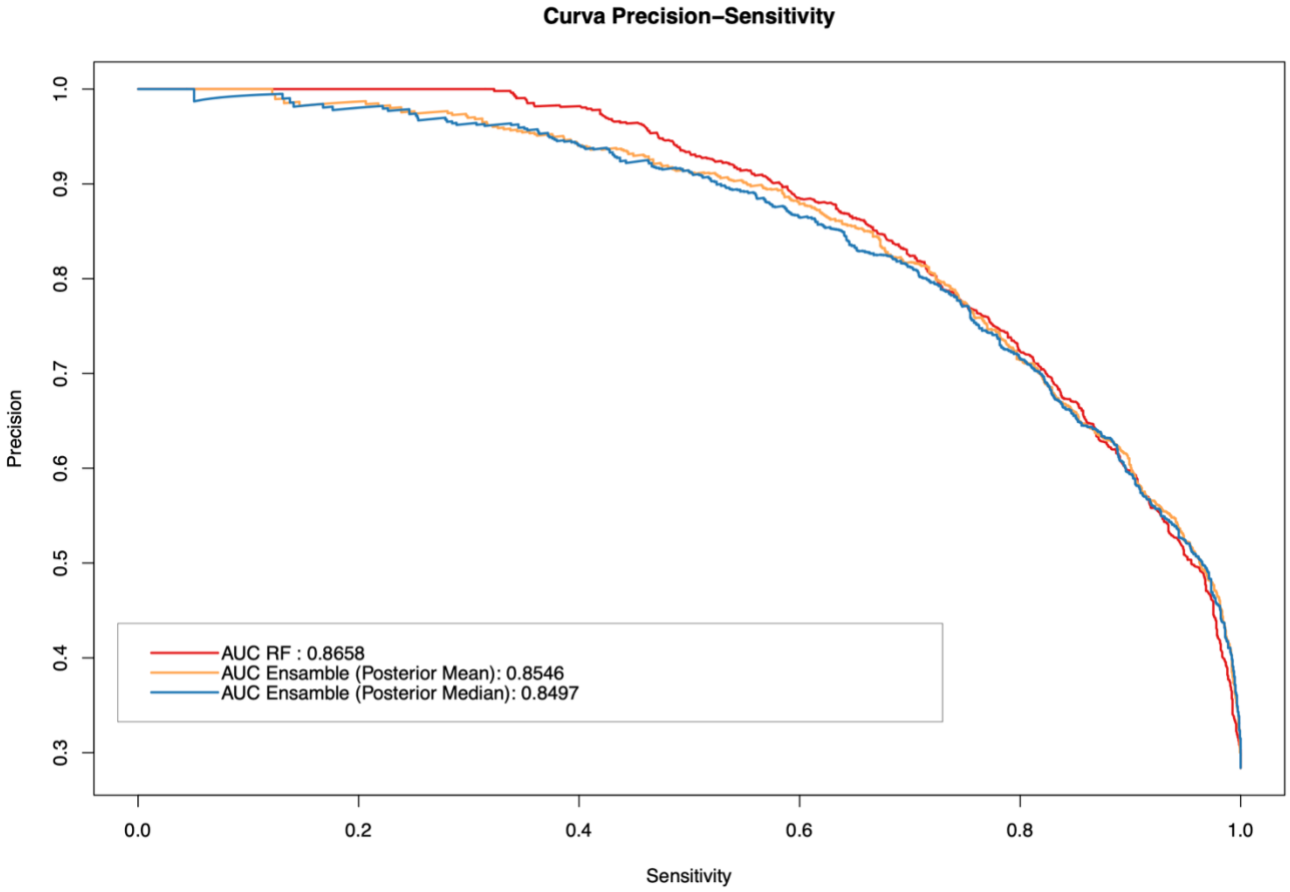
4.3 Model evaluation ed Ensamble methods

La metrica utilizzata per determinare il miglior modello in termini di performance è stata la curva PR. La **Precision-Recall curve** è uno strumento di valutazione utilizzato per analizzare le performance di un modello di classificazione. A differenza della curva ROC, che confronta la **Sensitility (Recall)** e la **Specificity**, la **curva PR** mette in relazione la **Precision** e la **Sensitivity** (**Recall**), le due metriche adottate per la stima dei modelli. La curva è ottenuta tracciando i valori di **Precision** sull'asse delle ordinate (y) e i valori di **Sensitivity (Recall)** sull'asse delle ascisse (x) per ogni possibile valore di soglia di classificazione del modello. **Una curva PR che si avvicina al punto (1,1)**, cioè nell'angolo in alto a destra del grafico per ogni singola soglia, indica che il modello ha alte prestazioni in termini sia di **Precision** che di **Sensitivity**. Il classificatore che sarà identificato come migliore è colui che presenterà l'area sottesa alla curva (**AUC**) massima.



La figura riporta le curve Precision-Recall (PR) per ciascun modello. Dall'analisi del grafico, si evince che le performance classificative dei predittori addestrati su dati bilanciati e sbilanciati sono simili, con i modelli addestrati su dati sbilanciati che mostrano delle performance superiori e una maggiore efficienza computazionale. In base a questi risultati, il modello che si è rivelato più performante è la Random Forest, con un'area sotto la curva (AUC) pari a 0,8658.

Infine, è stato eseguito un ensemble dei modelli classificativi, combinando le probability prediction di tutti i metodi, ad eccezione della Random Forest. L'obiettivo di questa combinazione è rafforzare le performance dei singoli modelli, al fine di ottenere previsioni più accurate e ridurne l'instabilità.



I risultati ottenuti mostrano che, complessivamente, l'insieme dei modelli ha prodotto performance competitive con quelle della Random Forest, sebbene quest'ultima rimanga la migliore.
Successivamente, per la Random Forest è stato stimato il valore di **cutoff** ideale per la classificazione di un cliente come churner. Attraverso un'ottimizzazione bayesiana 10 fold cross-validata, è emerso che la soglia in grado di massimizzare l'**F-Score** è pari a 0,3924.

In conclusione, il modello RF dagli iperparametri di **mtry**, **ntree** e **cutoff** ottimi, applicato sul test set, ha prodotto i seguenti risultati.

Random Forest Test Set				
F-Score	Precision	Sensitivity	Specificity	Accuracy
0,8026667	0,7984085	0,8069705	0,9194915	0,8876

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	301	76
Stayed	72	868

4.4 Variable Importance

Per valutare quali sono le variabili che hanno contribuito maggiormente nella discriminazione del target all'interno di ogni modello, è stata eseguita un'analisi della variable importance. L'importanza di ogni feature è riportatata nella seguente tabella.

Variable Importance	Random Forest	Albero di Classificazione	Modello Logistico	Neural Network	Radial Basis SVM	Mean Importance
Anzianita Mesi	100,00	77,87	73,33	100,00	100,00	90,24
Contratto: Two Year	35,98	57,38	94,41	72,45	98,48	71,74
Numero Referenze	44,27	100,00	100,00	56,22	53,81	70,86
Spese Totali	63,59	48,89	45,68	58,60	68,33	57,02
Numero Persone a Carico	15,21	45,98	68,10	99,34	42,12	54,15
Spese Totali Chiamate Lunga Distanza	32,68	32,58	0,00	58,16	66,60	38,00
Coniugato: Yes	5,08	7,48	92,91	27,81	33,66	33,39
Contratto: One Year	16,36	17,65	76,35	48,10	0,00	31,69
Spesa Mensile	32,51	53,61	0,00	33,34	36,68	31,23
Abbonamento Internet: Fiber Optic	16,08	55,70	5,32	53,67	0,00	26,15
Streaming Musica: Yes	2,13	0,00	0,00	74,77	15,00	18,38
Sicurezza Online: Yes	6,34	0,00	0,00	15,20	64,71	17,25

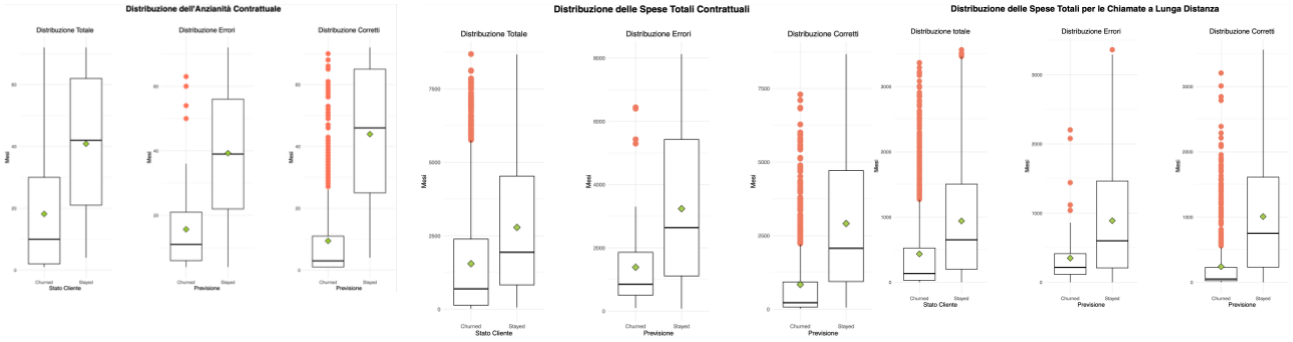
Analizzando i risultati è stato possibile osservare come le variabili **Anzianità in Mesi**, **Contratto**, (l'opzione biennale rispetto a quella mensile) e il **Numero di Referenze** sono, per tutti i modelli, le variabili più discriminanti. Inoltre, si evincono diverse caratteristiche, tra cui l'essere **Coniugati** e l'aver aderito al servizio di **Sicurezza Online**, che risultano fortemente significative, da un punto di vista classificativo, per solamente alcuni modelli. Tuttavia, uno dei principali limiti nell'utilizzo di **modelli black-box** è la difficoltà nell'interpretare il modo in cui le variabili predittive influenzano la classificazione degli individui. A differenza di modelli explainable come la **Regressione Logistica**, gli algoritmi più complessi, **Random Forest**, **Support Vector Machines** o **Reti Neurali**, non forniscono una chiara indicazione della direzione e dell'intensità dell'effetto di ciascun predittore sulla probabilità che un cliente receda dal proprio contratto. Dunque, non è possibile determinare in modo diretto quale sia il contributo specifico di ogni variabile nella decisione finale del modello individuato come ottimale. Nonostante questo, la **Regressione Logistica**, mediante i suoi i coefficienti stimati, permette di dare un'indicazione sull'influenza di ogni variabile sulla probabilità di churn. Nello specifico, applicando una trasformazione esponenziale ai coefficienti della regressione, è stato possibile interpretare il contributo delle variabili più importanti.

Variabile	Anzianità in Mesi	Contratto (Two Year)	Numero Referenze	Spese Totali	Numero persone a carico
Coefficiente	7,6986	3,4367	5,1512	0,2841	1,9031

I coefficienti riportati in tabella, ad eccezione delle **Spese Totali**, essendo maggiori di uno, evidenziano un impatto positivo sulla probabilità di recedere dal contratto. In particolare, ogni mese aggiuntivo di abbonamento (**Anzianità in mesi**) aumenta la probabilità di disdetta di sette volte rispetto al mese precedente, mentre la scelta di un contratto biennale, rispetto a uno mensile, la incrementa di tre. Questo potrebbe indicare una scarsa attività da parte della compagnia nella fidelizzazione del cliente che, dopo un periodo prolungato di sottoscrizione, decide di interrompere il servizio. Sorprendentemente, una spesa clientelare maggiore diminuisce la propensione al churn di circa il 70%. Questo potrebbe significare che i servizi proposti dall'azienda sono di alta qualità e che il cliente, che ricerca quel tipo di offerta, ne è estremamente soddisfatto e ben propenso alla spesa.

4.5 Anomaly Detection

In questa sezione, l'analisi si è focalizzata sugli errori di classificazione commessi dai modelli esaminati. In particolare, è stato effettuato un confronto tra le osservazioni difficili da classificare e quelle correttamente identificate, analizzando il comportamento dei soggetti rispetto le variabili che i modelli hanno identificato come le più rilevanti per la previsione.



Dall'analisi dei box plot emerge come le distribuzioni dei soggetti mal classificati siano risultate prossime a quelle dei soggetti correttamente predetti all'interno delle rispettive categorie. Questo suggerisce che gli errori non derivino dall'incapacità dei modelli di discriminare correttamente, ma piuttosto dal fatto che tali clienti risultino anomali rispetto al comportamento tipico delle loro classi di appartenenza. Inoltre, i grafici evidenziano il potere discriminante delle variabili, mostrando una chiara separazione tra le classi, a sottolineare il potere classificativo delle features selezionate nel processo.

4.6 Modelli basati sui costi

In modo analogo all'analisi effettuata sulle unità, per lo studio focalizzato sui profitti aziendali, al fine di massimizzarne i guadagni, è stato adottato un processo di addestramento mediante ottimizzazione bayesiana, con validazione incrociata 10-fold. In questo contesto, i classificatori adottati sono i medesimi ad eccezione della SVM in quanto non contente di adottare una ponderazione individuale per ogni singolo cliente. Gli iperparametri ottimi e le performance classificative stimate sono risultati le seguenti.

4.6.1 Modello Logistico

Entrata mensile totale stimata di 31.920 USD

Modello Logistico Costs			
F-Score	Precision	Sensitivity	Specificity
0,7370892	0,7395693	0,7346257	0,8975106

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1099	387
Stayed	397	3389

4.6.2 Albero di Classificazione

Entrata mensile totale stimata di 32.739 USD

Albero Classificativo Costs			
F-Score	Precision	Sensitivity	Specificity
0,727881	0,819933	0,6544118	0,9430614

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	979	215
Stayed	517	3561

4.6.3 Random Forest

Entrata mensile totale stimata di 33.677 USD

Random Forest Costs			
F-Score	Precision	Sensitivity	Specificity
0,743907	0,8471392	0,6631016	0,9525953

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	992	179
Stayed	504	3597

4.6.4 Neural Network

Entrata mensile totale stimata di 30.251 USD

Neural Network Costs			
F-Score	Precision	Sensitivity	Specificity
0,6958305	0,7118881	0,6804813	0,8908898

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	1018	412
Stayed	478	3364

4.7 Model evaluation basata sui costi

Analizzando i risultati è emerso che il modello in grado di massimizzare i profitti aziendali è la Random Forest con un entrata mensile totale stimata di 33.677 USD. Il classificatore, applicandolo al test set, ha prodotto le seguenti performance. La stima delle entrate medie totali previste sul test è pari a 87.707 USD.

RF basata sui costi

Random Forest Cost Test Set				
F-Score	Precision	Sensitivity	Specificity	Accuracy
0,7847731	0,8645161	0,7184987	0,9555085	0,8884

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	268	42
Stayed	105	902

RF basata sulle unità

Random Forest Test Set				
F-Score	Precision	Sensitivity	Specificity	Accuracy
0,8026667	0,7984085	0,8069705	0,9194915	0,8876

Confusion Matrix	Reference	
Prediction	Churned	Stayed
Churned	301	76
Stayed	72	868

Confrontando le stime ottenute con quelle della Random Forest addestrata tramite l'**F-Score**, si è osservato come le performance, sia in termini di costi che di unità, siano molto simili. In particolare, il classificatore allenato utilizzando la media armonica tra Sensitivity e Precision ha stimato **Entrate Mensili Totali** pari a 87.520 USD, contro le 87.707 USD ottenute dal modello basato sui costi.

Dal punto di vista dell'**F-Score**, la Cost Random Forest ha raggiunto un valore di 0.785, leggermente inferiore rispetto allo 0.803 della Random Forest in cui le unità non sono state ponderate. Tale risultato suggerisce che, in questo caso specifico, non vi fossero differenze significative nella spesa tra i clienti della compagnia, favorendo così una buona sovrapposizione nelle performance dei due classificatori.

5. Conclusioni e sviluppi futuri

Dall'analisi condotta, la Random Forest è risultata essere l'algoritmo più performante, sia in termini di accuratezza nella previsione del churn, che di gestione dei costi. L'integrazione di queste metodologie all'interno di un sistema aziendale potrebbe rivelarsi estremamente utile per individuare i clienti potenzialmente a rischio di disdetta, consentendo, non solo di attuare strategie mirate per trattenerli, ma anche di identificare gli aspetti contrattuali che necessitano di miglioramenti.

In particolare, l'adozione di un approccio basato sull'analisi costi-profitti permetterebbe un'allocazione più efficiente delle risorse aziendali, ottimizzando il budget destinato alle strategie di fidelizzazione.

Per quanto riguarda possibili sviluppi futuri, un'interessante estensione dell'analisi potrebbe riguardare la classificazione delle motivazioni alla base della disdetta e la segmentazione dei nuovi clienti rispetto agli storici. La prima consentirebbe di fornire alla compagnia informazioni utili per personalizzare le offerte di retention, riducendo significativamente il tasso di churn. La seconda, invece, permetterebbe di valutare l'efficacia delle campagne di marketing, verificando se il target raggiunto si sia effettivamente convertito in clienti attivi e fornendo così indicazioni strategiche per ottimizzare le future iniziative commerciali.

In conclusione, l'adozione di tecniche avanzate di machine learning, come la Random Forest, si dimostra un valido strumento per prevedere il churn e ottimizzare le strategie di retention aziendale.

Note a margine

L'overfitting negli AutoEncoder non rappresenta un problema in questo contesto, poiché consiste nella specializzazione del modello ai dati di training, limitandone l'applicabilità a nuove osservazioni. Tuttavia, se le osservazioni utilizzate per l'addestramento costituiscono l'intera popolazione disponibile, e non esistono altri dati al di fuori di questi, il concetto stesso di overfitting perde rilevanza.

Inoltre, lo scopo dell'AutoEncoder non è la classificazione, ma la riduzione non lineare della dimensionalità dei dati. Di conseguenza, la generalizzazione a nuove osservazioni non è un obiettivo, dato che non vi sono altri dati a cui applicare il modello.

L'overfitting diventerebbe problematico solo se il modello venisse addestrato su un sottoinsieme dei dati disponibili, poiché in tal caso la rete si specializzerebbe su quel campione senza catturare le caratteristiche generali dell'intera popolazione, compromettendo la qualità della riduzione dimensionale. Paradossalmente, in questo scenario, dove il campione coincide con l'intera popolazione, l'obiettivo è proprio ottenere un modello altamente specializzato, capace di apprendere perfettamente la struttura dei dati per ridurne la dimensionalità preservandone il massimo delle informazioni.