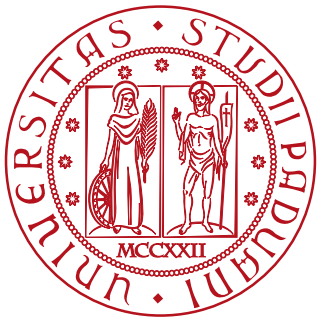
La vendita incrociata di prodotti: il caso di una Banca di Credito Cooperativo

Mattia Compagno, Riccardo Fassina, Matteo Gasparin

Strumenti Statistici per l'Analisi di Dati Aziendali

A.A. 2021-2022

Università degli Studi di Padova

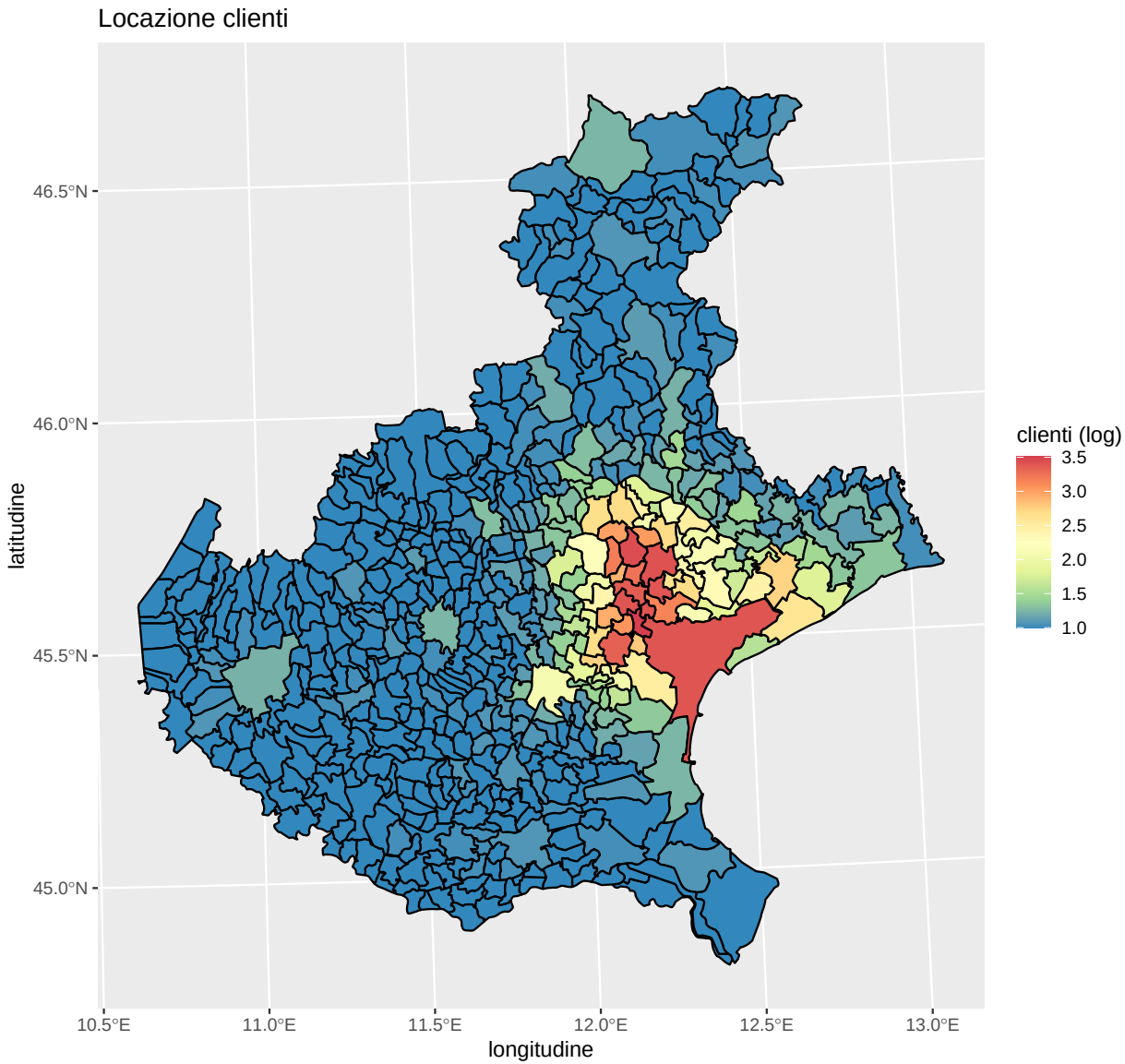


Introduzione

Dataset

Il dataset analizzato proviene dall'unione di diversi *data mart* riguardanti la clientela di una Banca di Credito Cooperativo operante in Veneto. In particolare sono presenti informazioni riguardanti 27 variabili inerenti a 39 600 clienti rilevate nell'anno 2018.

BCC


Locazione clienti
46.5°N
46.0°N
45.5°N
45.0°N
latitudine
10.5°E
11.0°E
11.5°E
12.0°E
12.5°E
13.0°E
longitudine
clienti (log)
3.5
3.0
2.5
2.0
1.5
1.0

Variabili

In particolare, il dataset, contiene:

- variabili anagrafiche;
- variabile relative al Conto Corrente;
- variabili riguardanti il rapporto tra Banca e cliente;
- ulteriori variabili.

Analisi preliminare

Per quanto concerne l'analisi preliminare si è proceduto ad eliminare variabili ritenute superflue per l'analisi condotta, o comunque che fossero ridondanti. Si è successivamente deciso di accorpare le modalità di alcune variabili. Il dataset ottenuto è così composto da 39 510 clienti e 19 variabili.

Analisi preliminare

Variabile	Descrizione
Codice_filiale_cliente	Filiale di appartenenza del cliente
Comune_residenza_o_sede	Comune residenza del cliente
Eta	Età
Grado_di_rischio	Indicatrice del grado di rischio del cliente
Mesi_anzianita_cliente	Età (in mesi) del rapporto tra la Banca ed il cliente
Professione	Professione cliente
Sottosettore_attivita_economica	Codice ATECO cliente
ImpieghiSMLa	Saldo medio liquido degli impieghi
RaccDirSMLa	Saldo medio liquido della raccolta diretta
RaccIndSMLa	Saldo medio liquido della raccolta indiretta
Mfa	Margine finanziario
Msa	Margine servizi
Dist	Distanza del cliente dalla filiale
Sport	Numero operazioni effettuate allo sportello
Self	Numero operazioni effettuate al self-service
Tipo_controparte2	Tipologia di cliente
Area	Provincia della filiale
p_18	Numero prodotti posseduti nel 2018
p_19	Numero prodotti posseduti nel 2019

Analisi esplorative

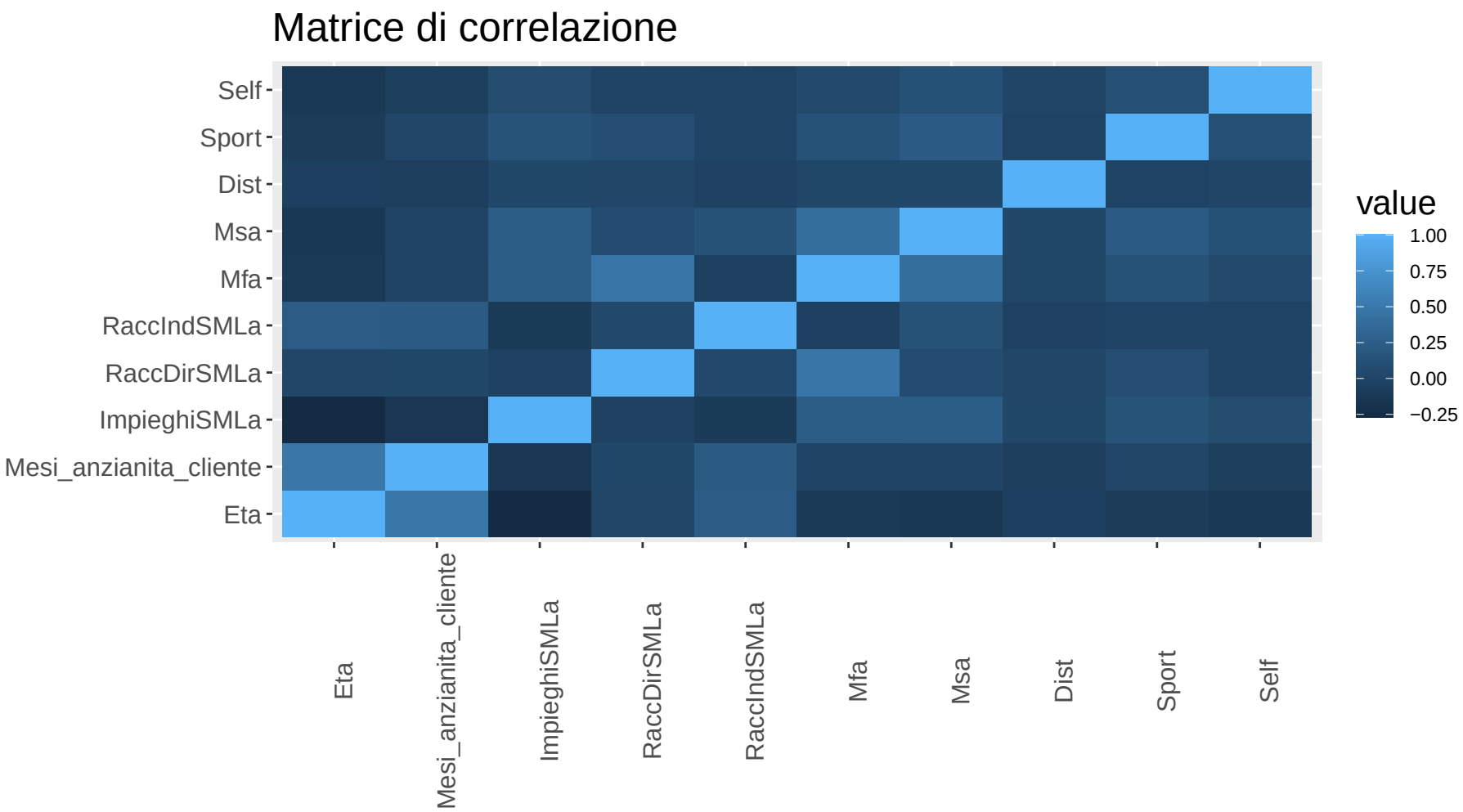


Figura: Correlazioni tra variabili quantitative

Analisi esplorative

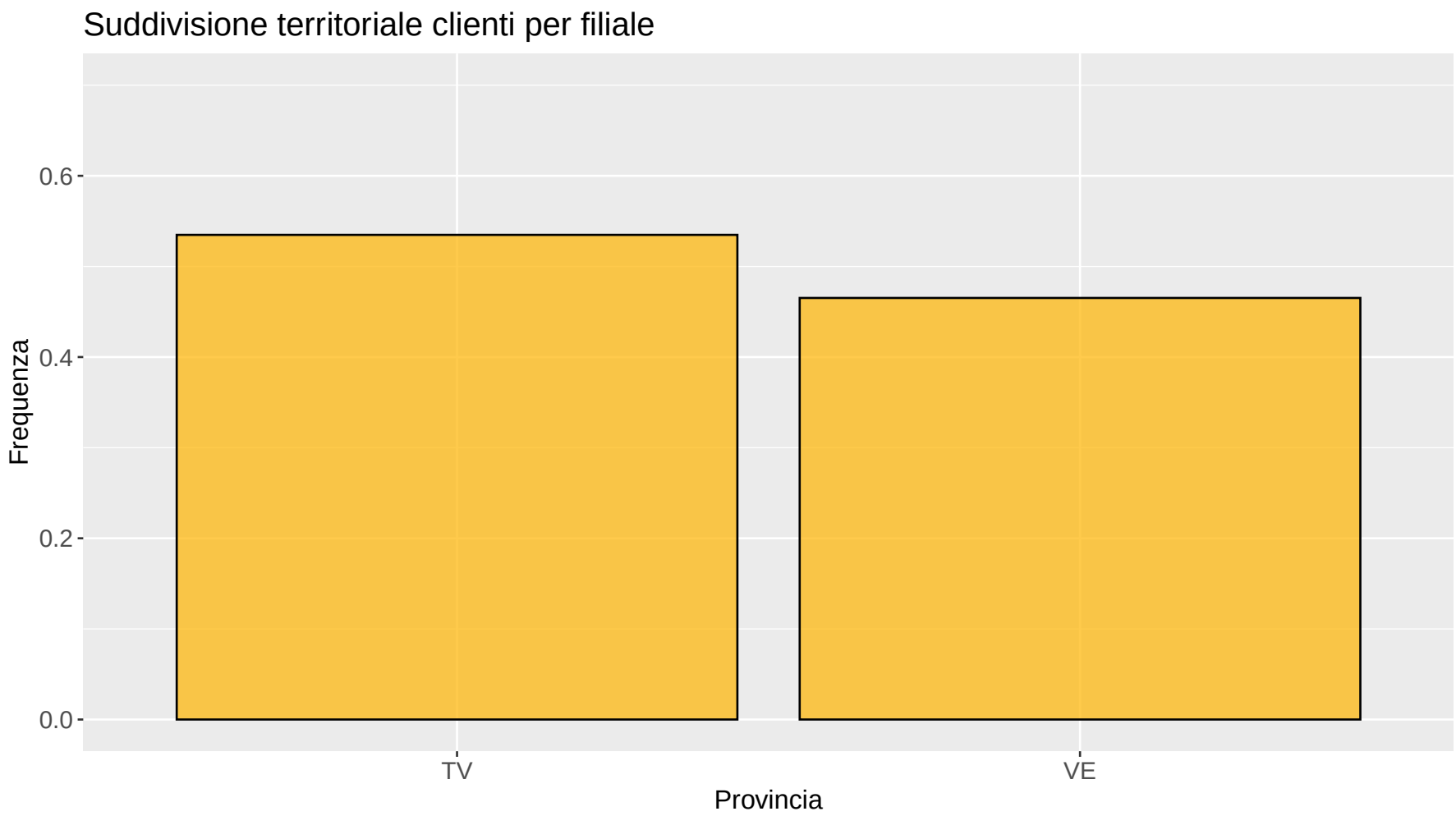


Figura: Suddivisione clienti tra filiali di TV e VE

Analisi esplorative

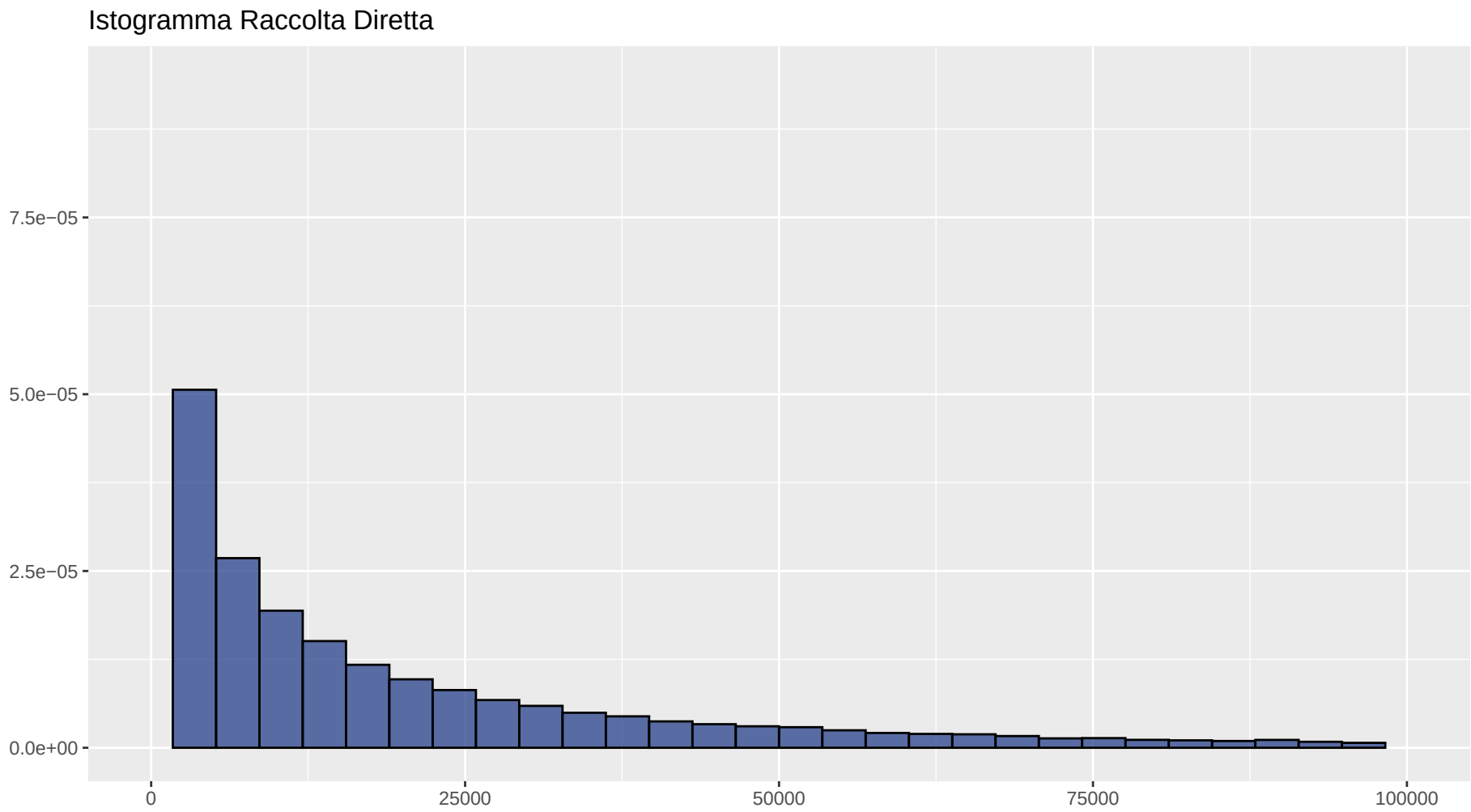


Figura: Istogramma raccolta diretta

Prima domanda di business

Il contesto

Il numero di prodotti posseduti da un cliente è una variabile di fondamentale importanza per un istituto bancario. Tale importanza deriva da due principali motivi:

1. le commissioni sui prodotti posseduti dal cliente sono un importante fonte di guadagno per la Banca;
2. il possesso di più prodotti riduce il rischio di abbandono da parte della banca.

Churn Rate

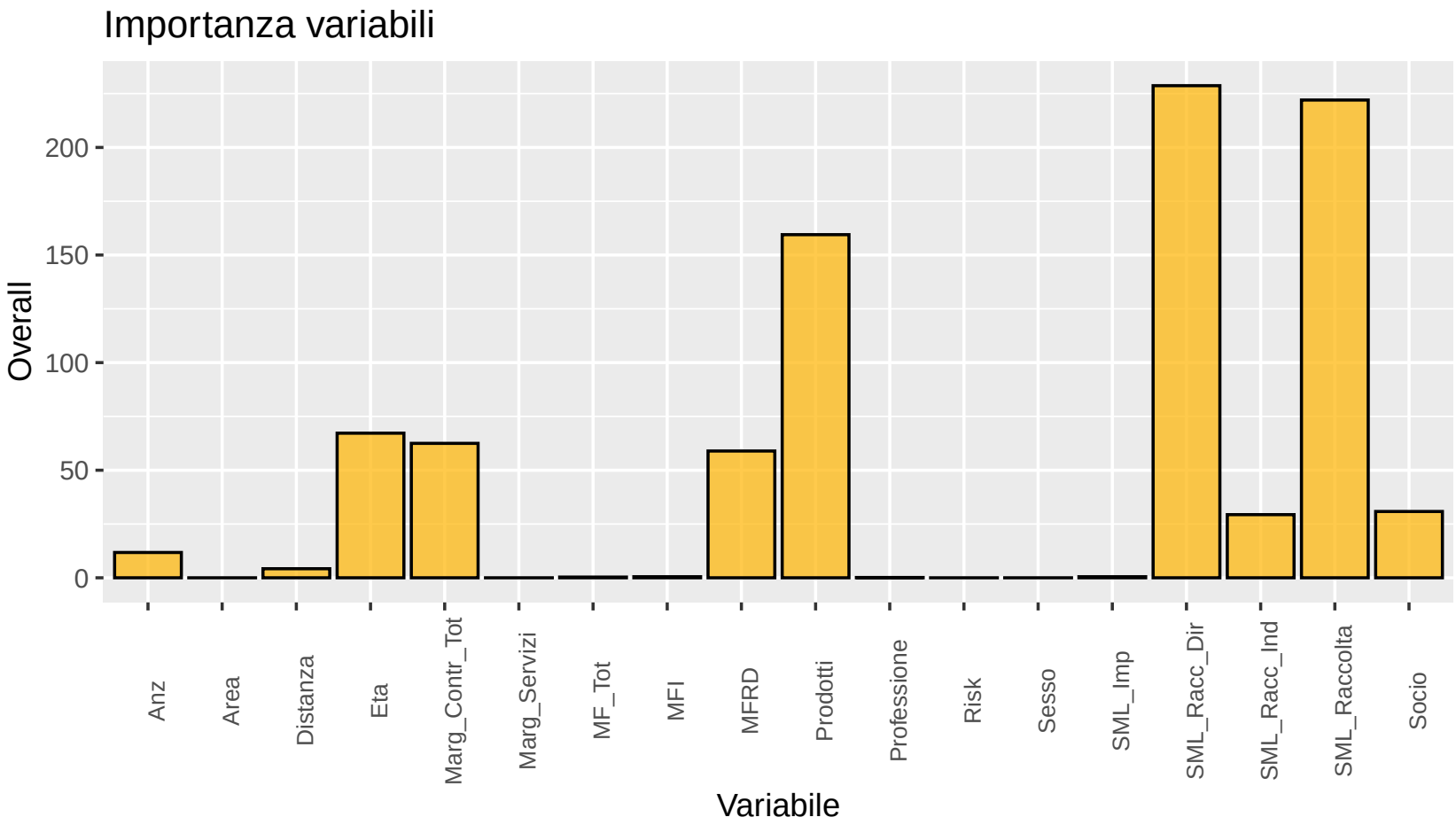


Figura: Importanza variabili calcolata tramite caduta di devianza effettuata in una precedente analisi riguardante il Churn

Riduzione del Churn

Vista l'elevata concorrenza tra i vari gruppi bancari, si ha che il tasso di abbandono è un indice di vitale importanza. In particolare, tale misura assume ancora più spessore per una banca di medio-piccole dimensioni che non sempre riesce a competere con i grandi colossi del settore. Operazioni di *cross-selling* o *up-selling* sono dunque di fondamentale rilevanza per aumentare la fedeltà della clientela.
Risulta quindi utile andare a considerare il numero di prodotti posseduti dal cliente per prendere decisioni tempestive atte a prevenire tale rischio.

Variabile risposta

Si considera quindi una regressione sulla differenza prodotti tra un anno (2018) ed il successivo (2019). Dal calcolo di questa quantità si nota la presenza di un numero elevato di zeri (circa il 54% del totale).
Una prima scelta potrebbe consistere nel considerare come variabile risposta

$$y = (p_{19} - p_{18})$$

Tale formulazione però è definita in un supporto discreto ed inoltre non tiene conto dell'entità relativa del cambiamento.

Variabile risposta

Si è quindi deciso di lavorare in scala percentuale

$$y = \frac{p_{19} - p_{18}}{p_{18}}.$$

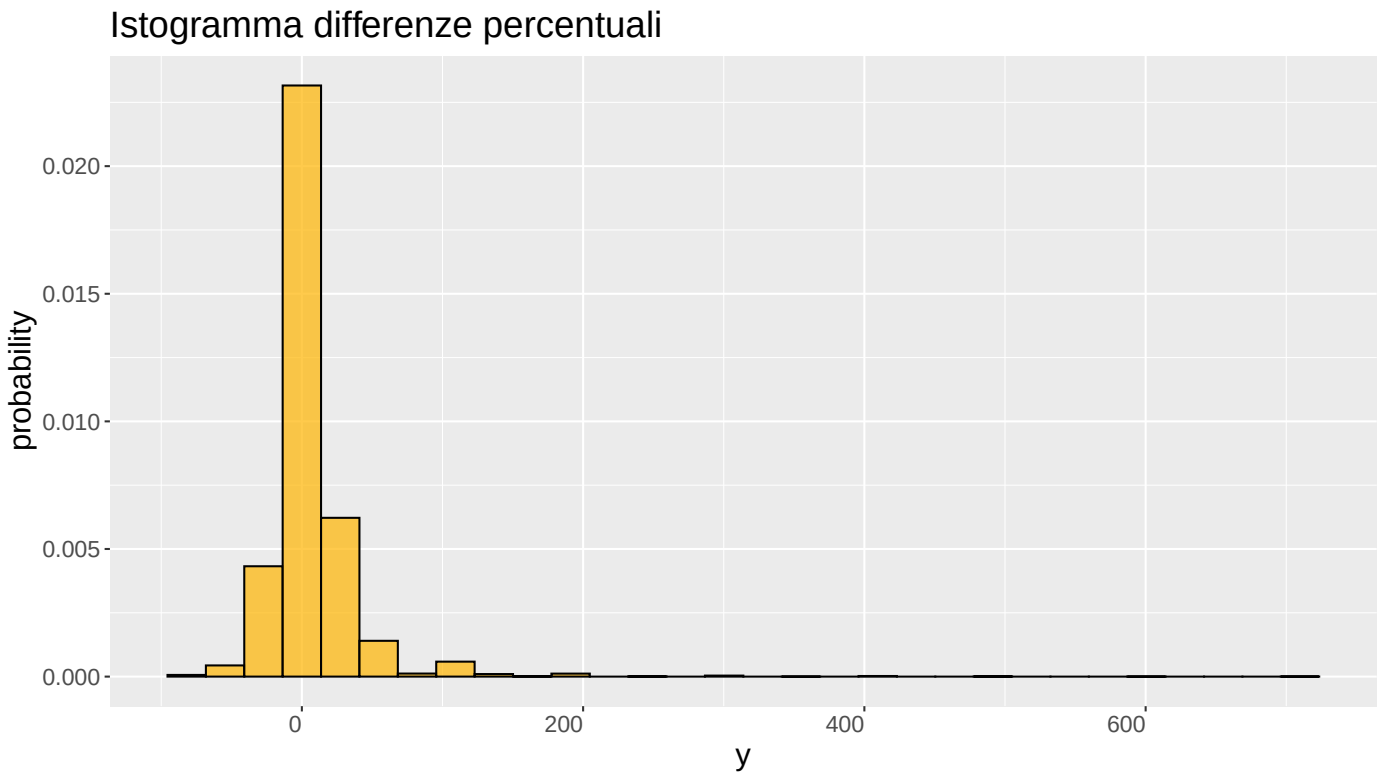


Figura: Istogramma differenze percentuali

Variabile risposta

Al fine di rendere simmetrica la distribuzione marginale, si è lavorato in scala logaritmica

$$y = \log(p_{19}) - \log(p_{18}),$$

che di fatto è uno sviluppo in serie di *McLaurin* della quantità precedente.

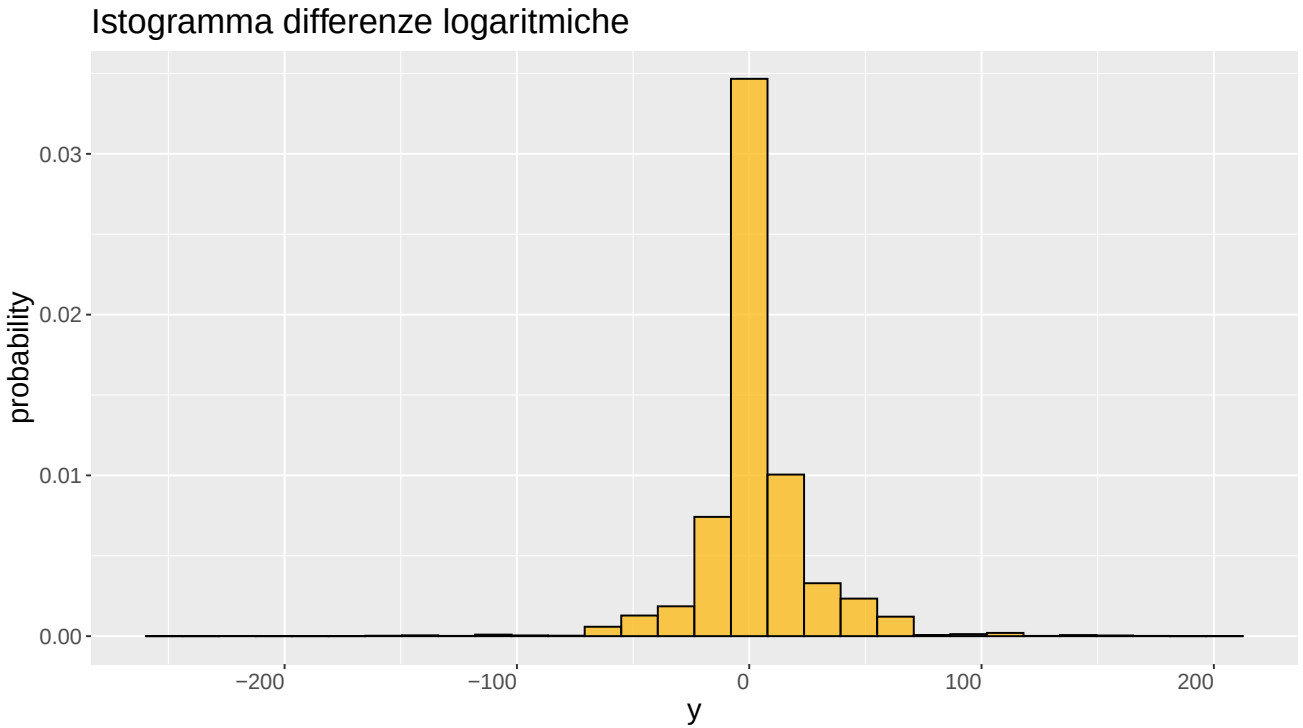


Figura: Istogramma differenze logaritmiche

Suddivisione dataset

Visto l'elevato numero di osservazioni si è proceduto con la seguente suddivisione del dataset:

- insieme di stima $\rightarrow$ 50%;
- insieme di convalida $\rightarrow$ 25%;
- insieme di verifica $\rightarrow$ 25%.

Modellazione

Al fine di prevedere e spiegare la risposta si sono adattati vari modelli di regressione. La metrica da noi proposta al fine di tenere conto sia dell'entità sia della natura della risposta, è la seguente

$$MSEC = \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{y}_i - y_i)^2 + w\ \mathbb{1}_{(\hat{y}_i y_i<0)}\ \hat{y}_i{}^2\right],$$

dove $w \geq 0$ è un peso che penalizza l'errata attribuzione del segno. Si noti che se $w = 0$ si ha l'usuale metrica MSE.

Risultati

	MSE	MSEC ($w = 0.25$)	MSEC ($w = 0.5$)	MSEC ($w = 1$)
Naive [1]	588.46	588.46	588.46	588.46
Lineare	569.15	570.08	571.02	572.88
Step	569.52	570.44	571.35	573.19
Ridge	568.89	569.68	570.48	572.06
Lasso	569.28	570.08	570.88	572.48
Relaxed lasso	569.41	570.32	571.24	573.07
GAM	564.14	565.24	566.33	568.52
Albero	548.28	549.34	550.41	552.54
MARS	535.12	536.42	537.72	540.32
Rete Neurale	547.35	548.70	550.05	552.76
GBM	517.81	518.86	519.91	522.02

[1] Modello che prevede $\hat{y}_i = 0 \ \forall i$ considerata l'elevata presenza di zeri.

Interpretazione - GAM

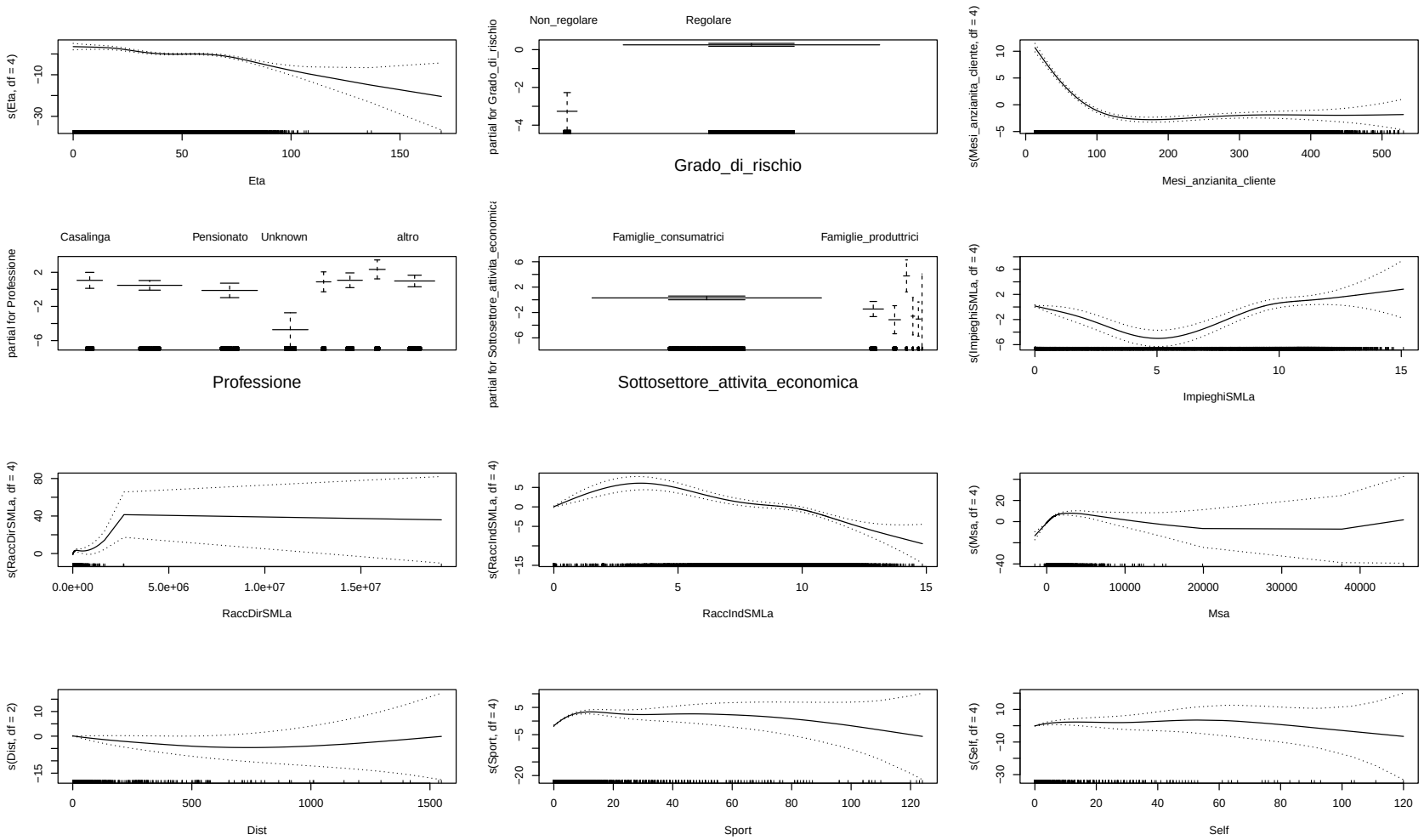


Figura: Relazioni marginali GAM

Interpretazione - GBM

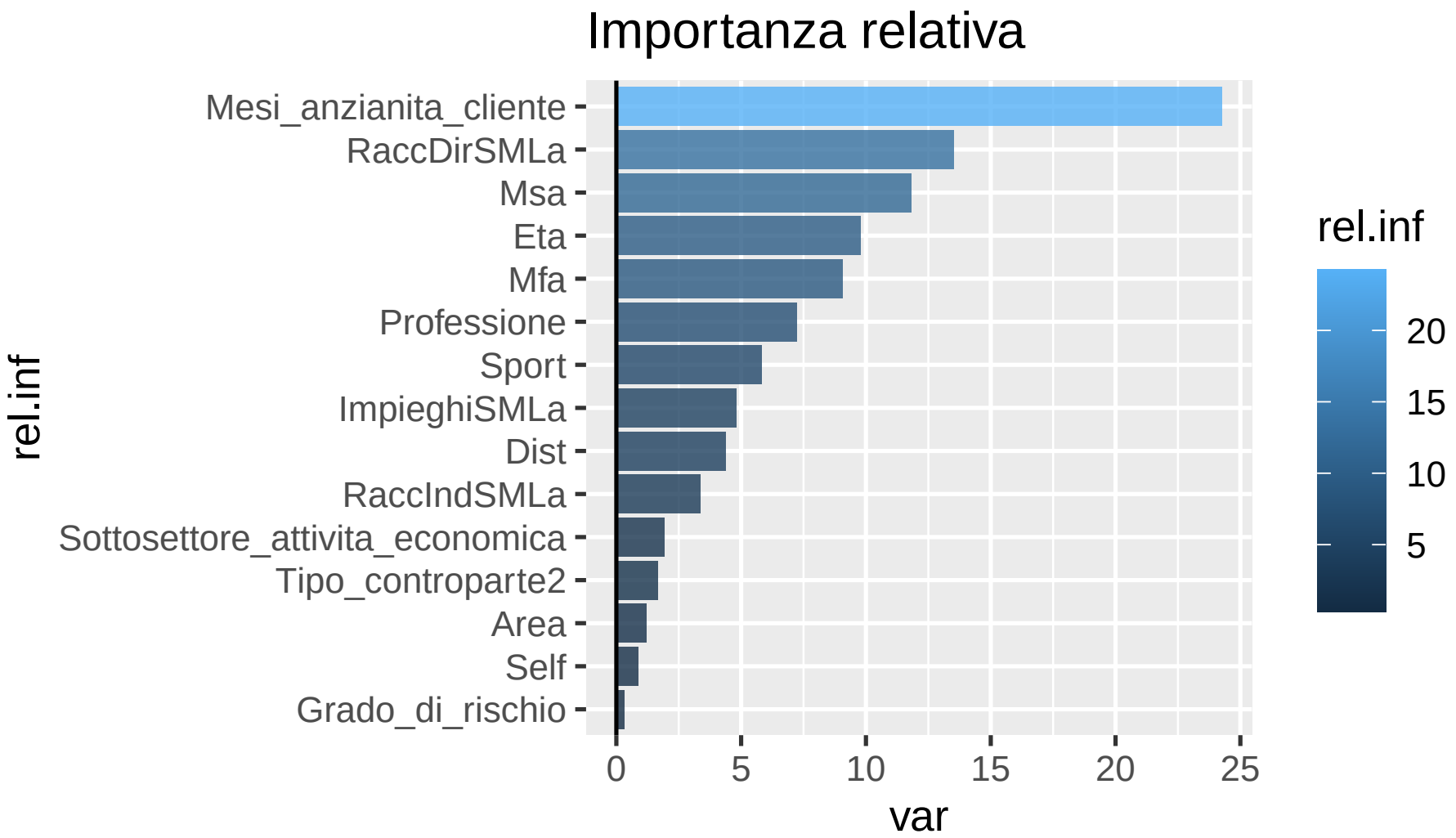


Figura: Importanza variabili GBM

Interpretazione - GBM

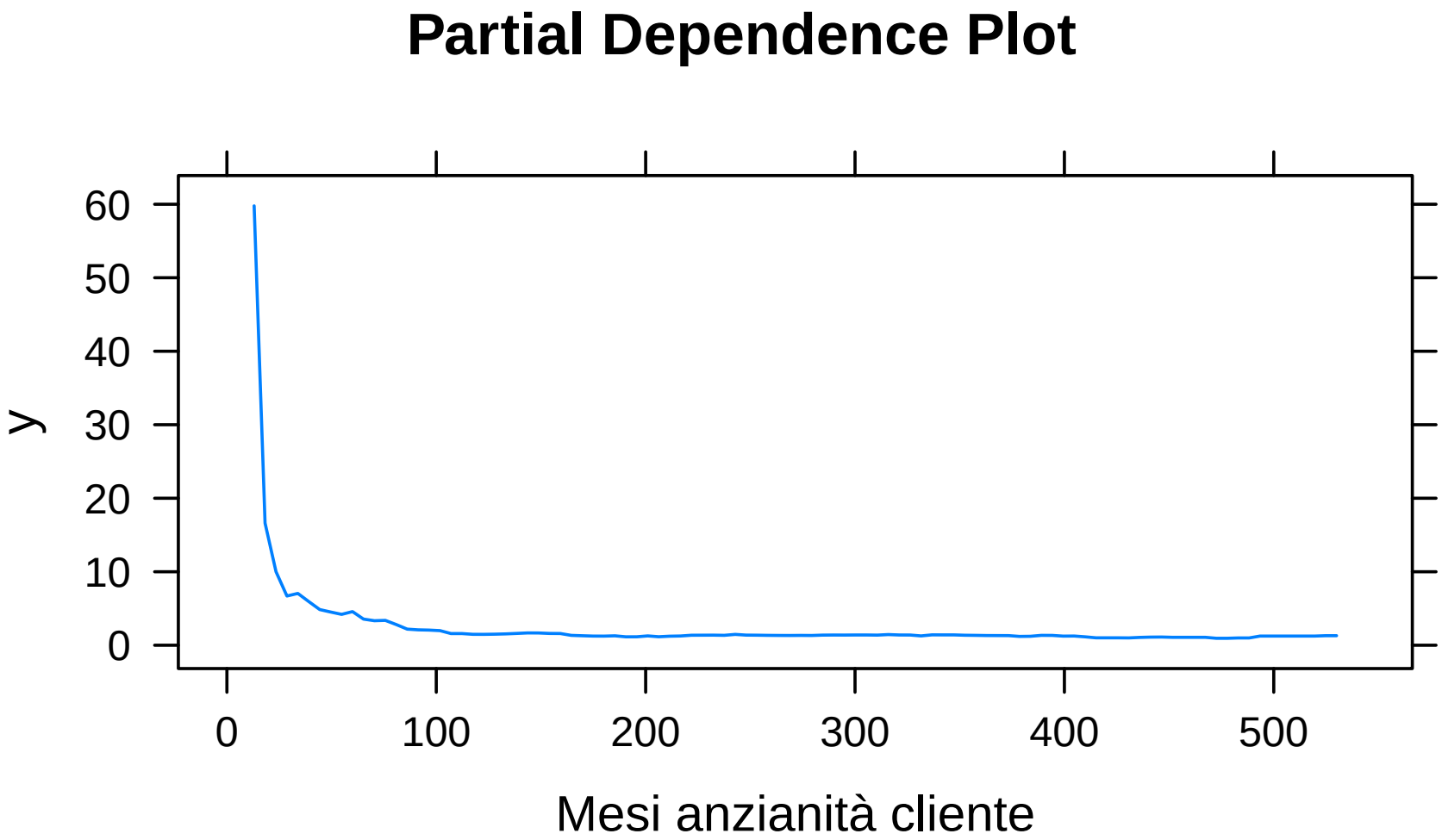


Figura: Partial dependence plot

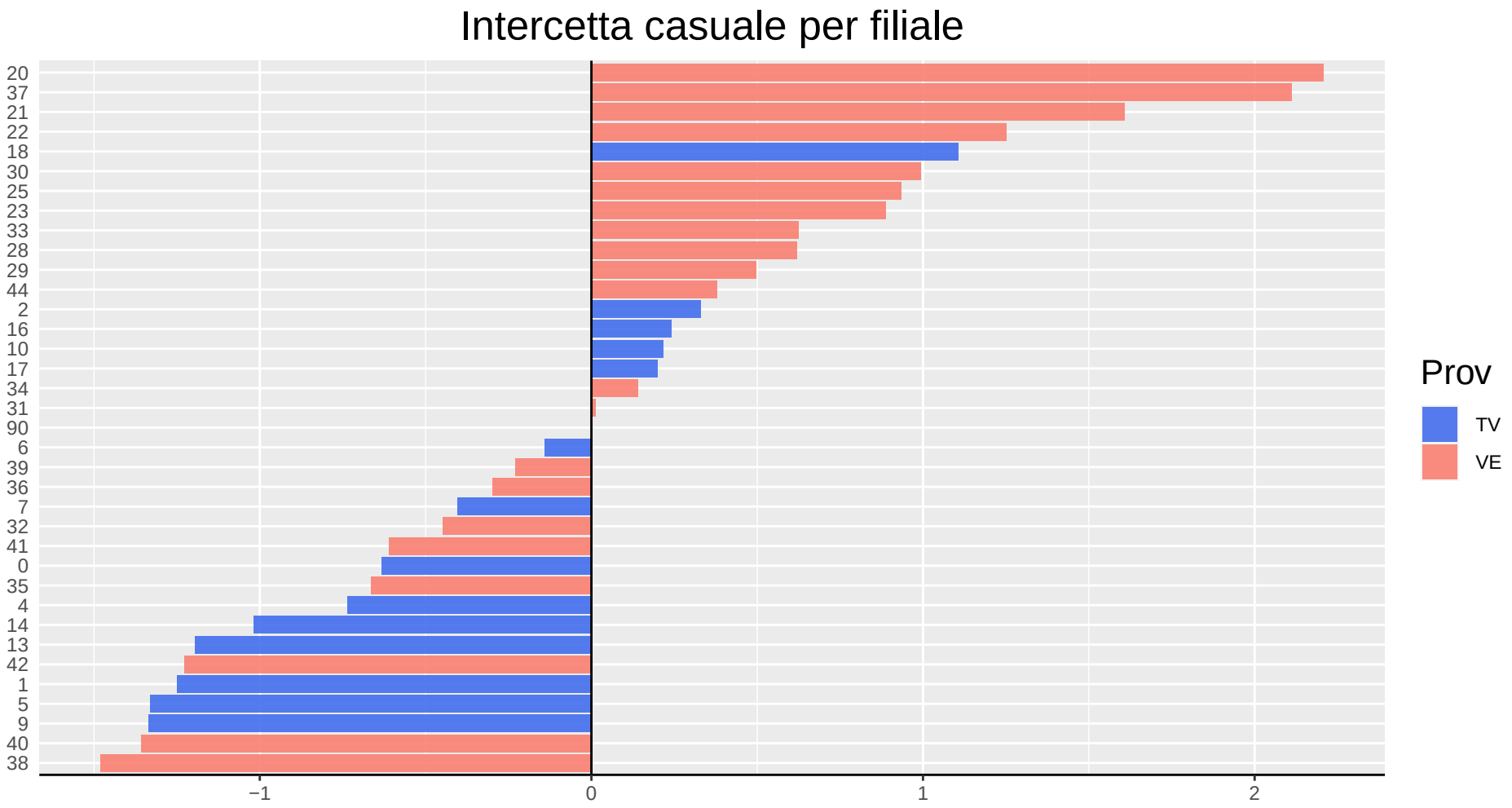
Modello gerarchico

Non avendo considerato nei precedenti modelli la variabile `Codice_filiale_cliente`, si è considerato opportuno introdurla con un modello gerarchico ad intercetta casuale, specificato come segue:

$$y_i = \alpha_{j[i]} + \beta^\top x_i + \epsilon_i,$$

dove $\alpha_{j[i]}$ è il valore assunto dall'intercetta casuale per la j-esima filiale. Le covariate utilizzate sono quelle selezionate dal modello stepwise e l'MSE conseguito nell'insieme di verifica risulta pari a 569.03, molto simile a quello dei modelli lineari.

Interpretazione - Modello gerarchico



Modellazione in due stadi

Si è cercato di risolvere il problema dell'inflazione di zeri, in tal senso si è quindi proceduto ad una modellazione in due stadi. Si presume che

$$Y_i = \begin{cases} 0 & \text{con probabilità} = p_i, \\ Y_i^* & \text{con probabilità} = 1 - p_i. \end{cases}$$

Si procede quindi come segue

1. modellare la probabilità che la risposta Y_i sia pari a 0;
2. modellare Y_i^* tramite una regressione sugli $Y_i \neq 0$.

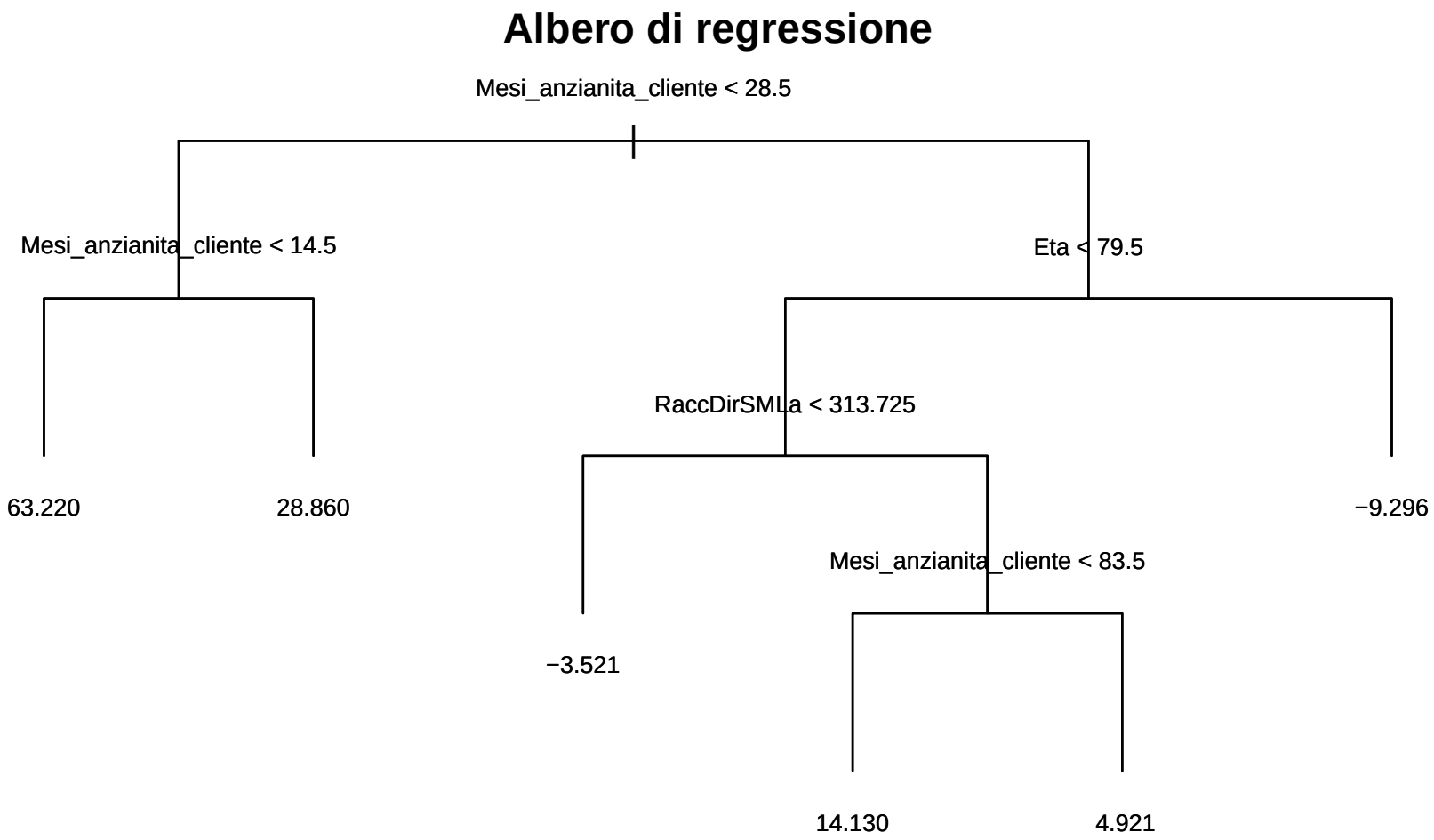
La previsione su una nuova unità sarà quindi $\hat{y}_i = (1 - \hat{p}_i) \cdot \hat{y}_i{}^*$.

Modellazione in due stadi - Risultati

Il primo passo ha visto l'utilizzo di un modello di regressione logistica stepwise, mentre per il secondo passo si sono provati diversi modelli. I risultati ottenuti sono i seguenti.

	MSE	MSEC ($w = 0.25$)	MSEC ($w = 0.5$)	MSEC ($w = 1$)
Lineare	570.28	571.34	572.40	574.52
Albero	552.64	553.56	554.47	556.31
MARS	577.71	578.76	579.81	581.91
GBM	553.18	553.68	554.17	555.17

Albero di regressione



Commenti

La modellazione a due stadi peggiora i risultati precedenti, in quanto il modello non è in grado di distinguere adeguatamente le due popolazioni. Questo perchè le due popolazioni non sono tra loro nettamente distinguibili.
Nel complesso, il modello migliore è il *GBM* relativo alla prima modellazione, il quale presenta un MSE pari a 517.81 ed un tasso di errata classificazione del 34%: il più basso tra tutti i modelli.

Seconda domanda di business

Segmentazione

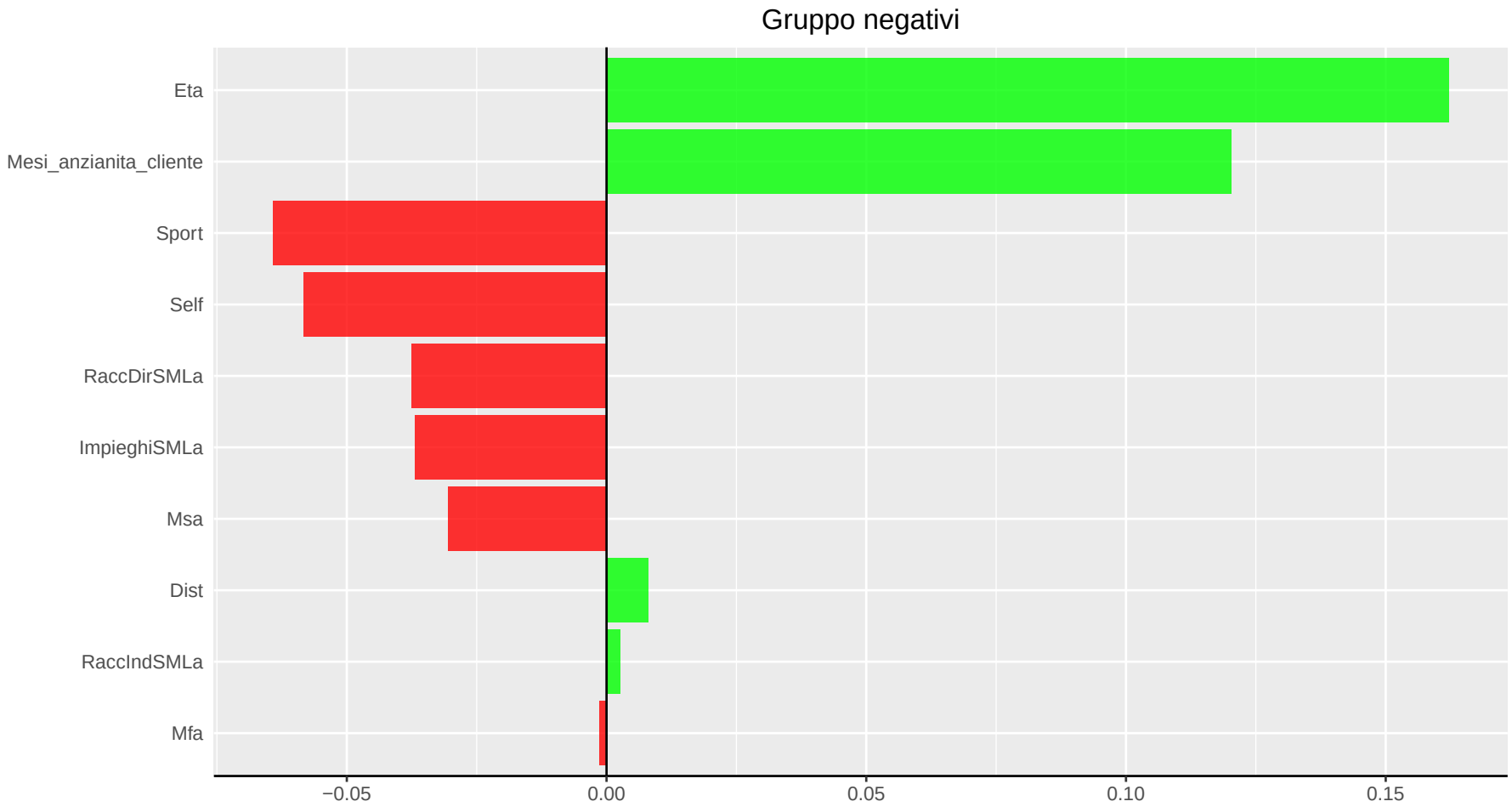
Risulta ora d'interesse per l'istituto bancario andare a caratterizzare i clienti che perdono prodotti durante l'anno. Dato che le aziende presentano delle dinamiche differenti dalle persone fisiche si è preferito escluderle durante queste analisi.

Scopo dell'analisi

Si effettua quindi un clustering al fine di scovare dei gruppi di unità simili con lo scopo di proporre delle soluzioni *ad hoc* per ciascun gruppo. Riuscire a personalizzare il più possibile le offerte può risultare cruciale per la fidelizzazione del cliente.

Analisi esplorativa

Come primo passo si sono raffrontate le caratteristiche dei clienti che hanno acquisito o perso prodotti.



Riduzione della dimensionalità

Ci si è concentrati quindi sui clienti che hanno perso prodotti.
Dato che molti metodi di clustering si basano su misure di distanza sullo spazio delle variabili esplicative, si è preferito escludere le variabili qualitative nella definizione dei vari gruppi mantenendo così solamente le quantitative (standardizzate). L'applicazione dei metodi di raggruppamento sui dati originali non ha prodotto buoni risultati, probabilmente vista l'elevata asimmetria di variabili relative al denaro. Si è pensato di ricorrere ad un metodo di riduzione della dimensionalità al fine di effettuare un miglior clustering.

t-SNE

Il metodo da noi utilizzato a tal fine è il *t*-SNE (t-distributed Stochastic Neighbor Embedding). Tale metodologia cerca di ridurre la dimensionalità del problema cercando di mantenere le distanze originali tra le osservazioni.

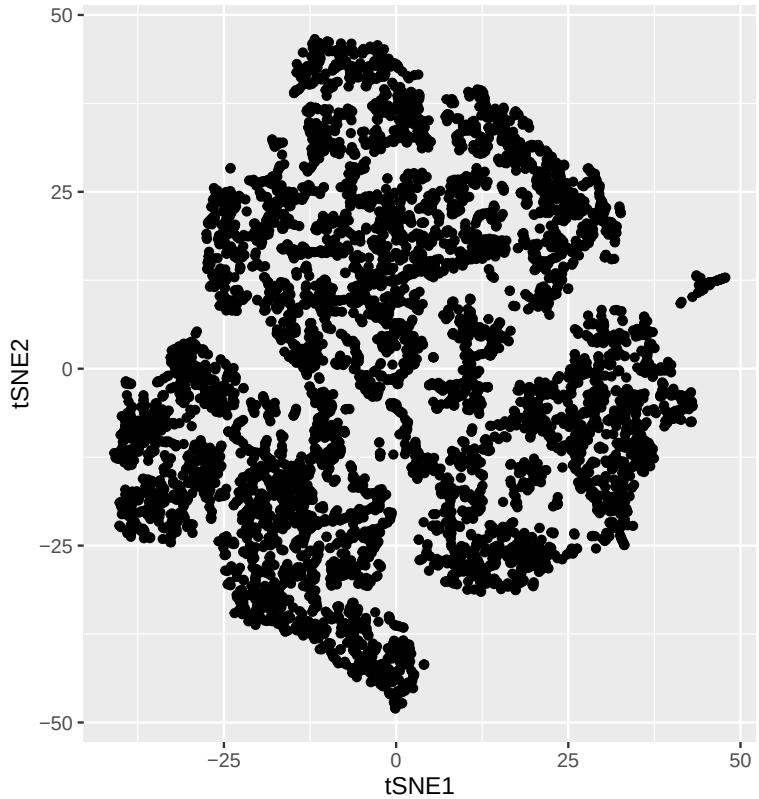


Figura: Dimensioni t-SNE

Clustering gerarchico

Il primo metodo utilizzato è il clustering gerarchico, dopo varie configurazioni si è deciso di utilizzare il legame *medio* sulla matrice contenente le distanze *euclidee* tra i punti. Per l'interpretabilità si è adottato un numero di cluster contenuto, in questo caso $K = 3$.

Clustering gerarchico

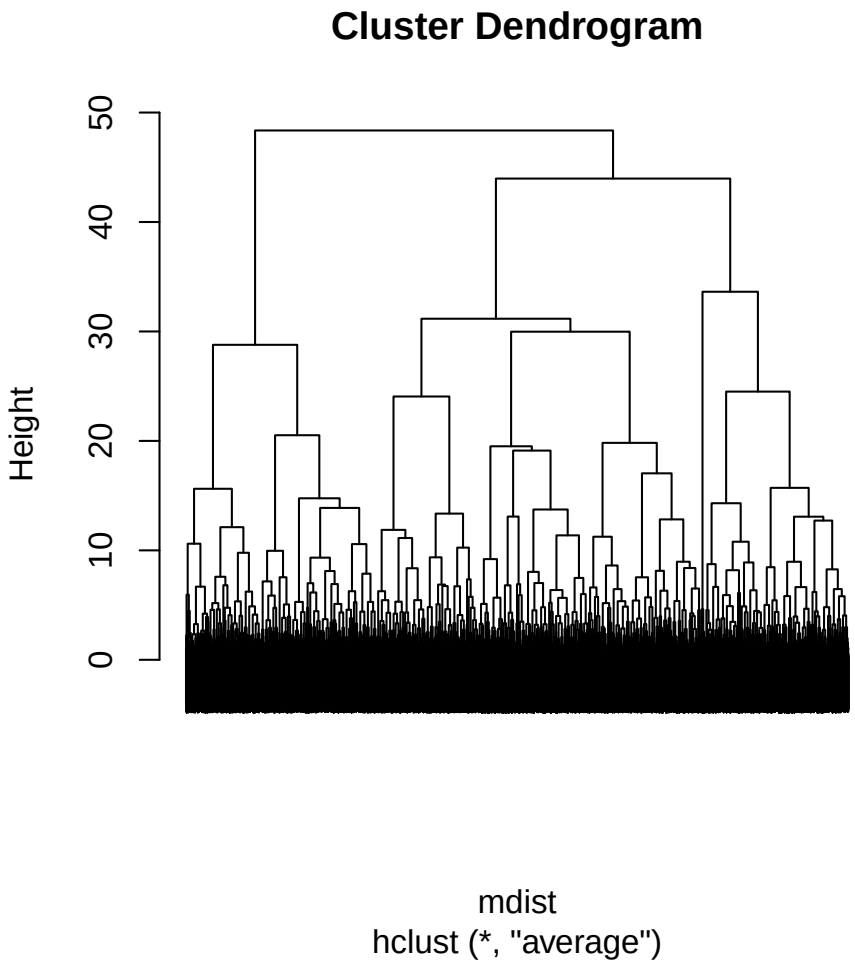
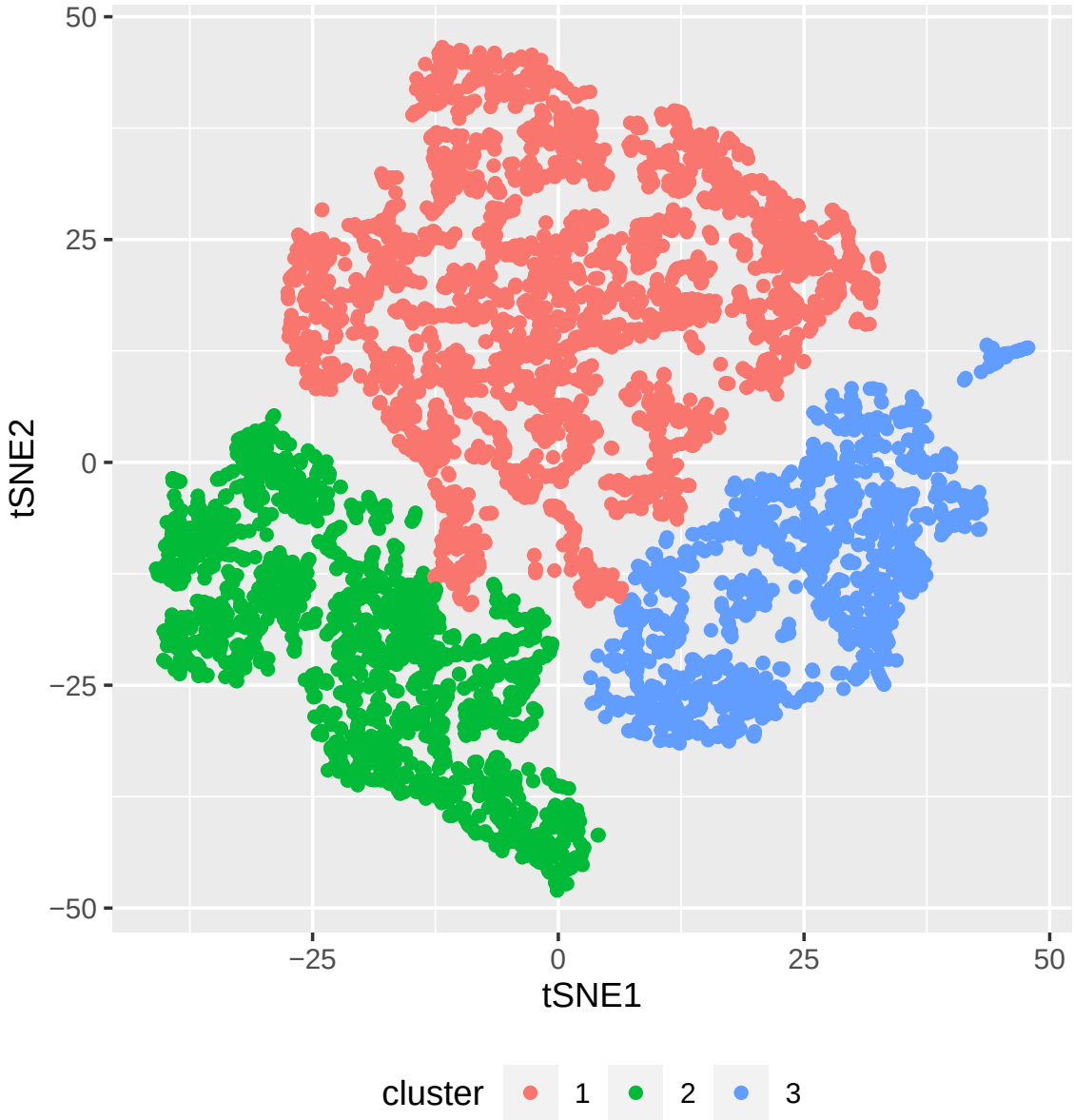


Figura: Dendrogramma clustering gerarchico agglomerativo

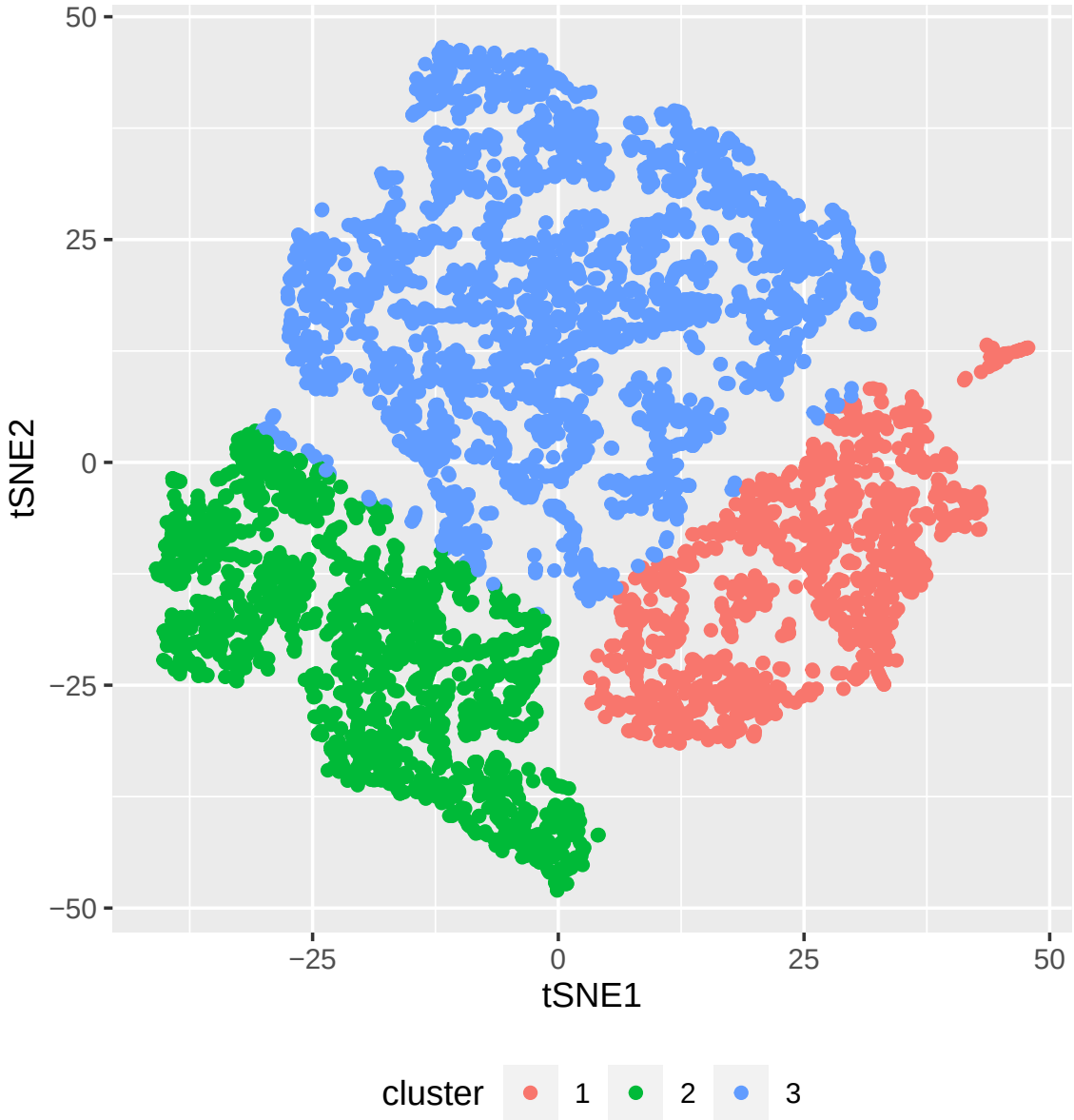
Clustering gerarchico



Model-based clustering

Succesivamente si è provato ad adottare una metodologia model-based. Il parametro di regolazione da scegliere è il numero di componenti facenti parti del modello mistura. Dato che il BIC sceglieva un numero di componenti molto elevato, si è deciso a fini interpretativi di scegliere comunque $K = 3$. La struttura di Σ_k selezionata dal BIC è *VVE* (distribuzione: ellissoidale, volume: variabile, forma: variabile, orientamento: uguale)

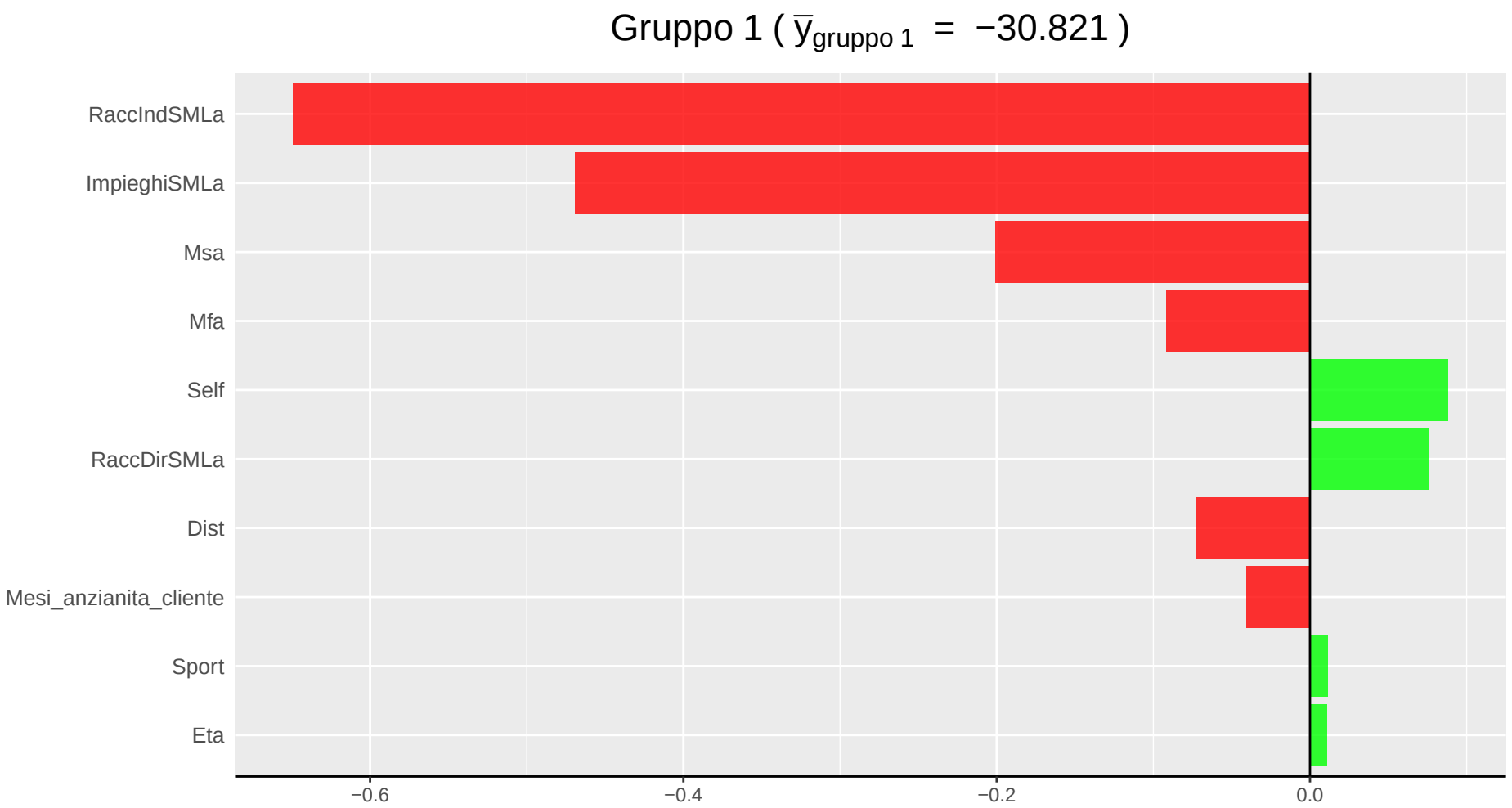
Model-based clustering



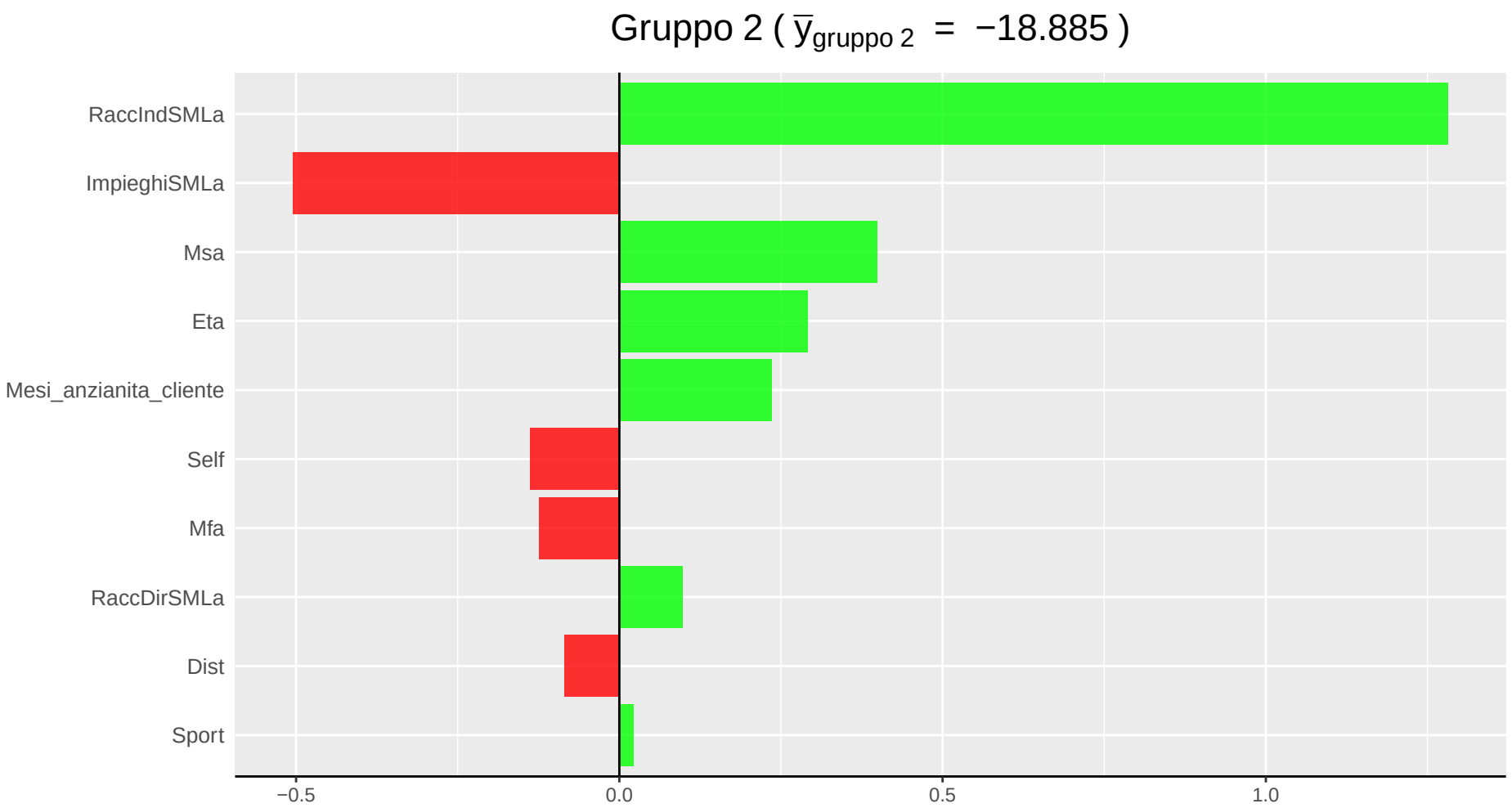
Commenti

Come si nota dalle figure, i gruppi selezionati sembrano meglio selezionati dal clustering gerarchico in quanto sembra ci sia una maggior distanza tra i gruppi. Si procede quindi alla caratterizzazione dei vari gruppi.

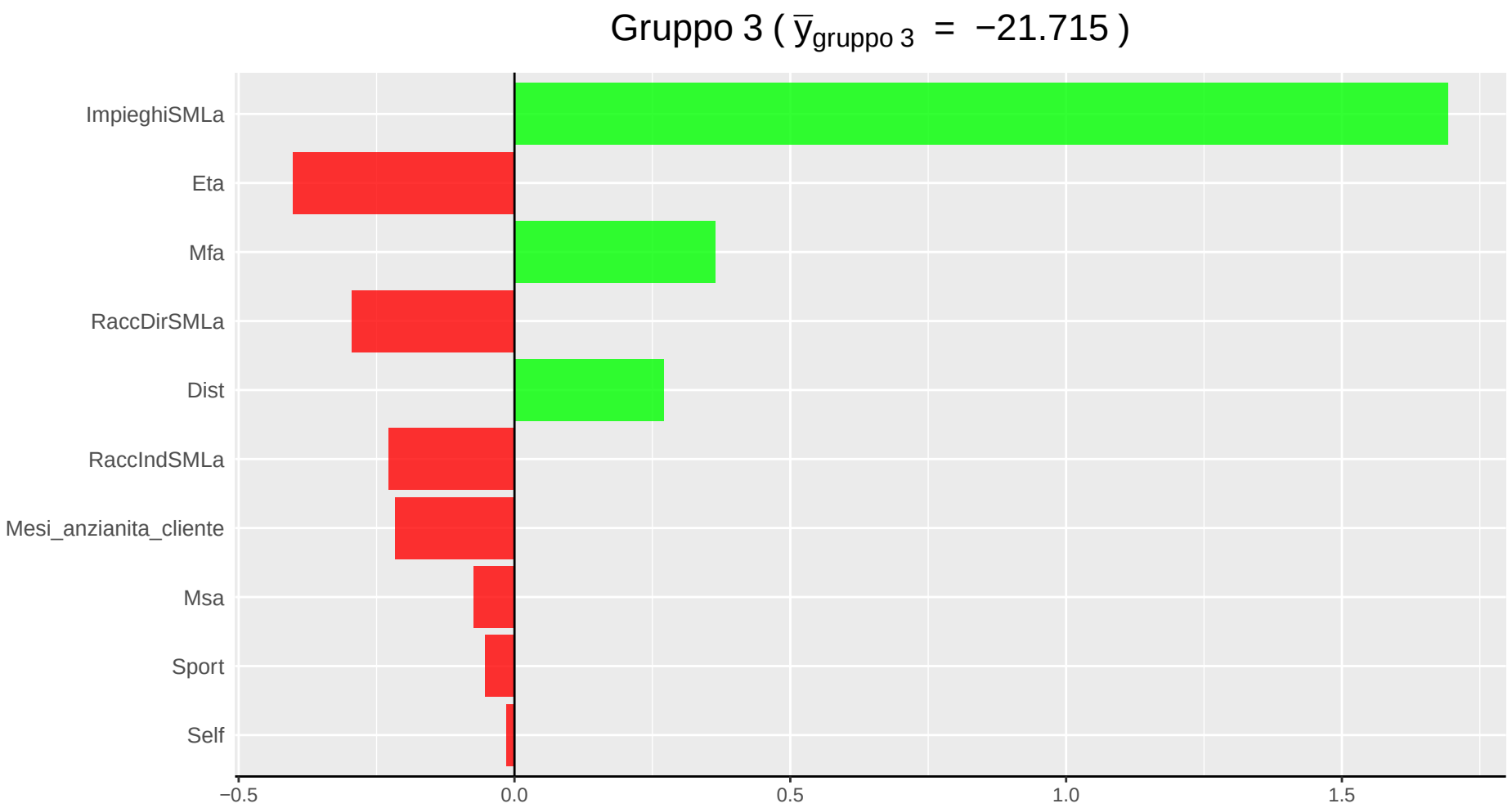
Gruppo 1


Gruppo 1 ($\overline{y}_{\text{gruppo 1}} = -30.821$)
RaccIndSMLa
ImpieghiSMLa
Msa
Mfa
Self
RaccDirSMLa
Dist
Mesi_anzianita_cliente
Sport
Eta
−0.6
−0.4
−0.2
0.0

Gruppo 2


Gruppo 2 ($\overline{y}_{\text{gruppo 2}} = -18.885$)
RaccIndSMLa
ImpieghiSMLa
Msa
Eta
Mesi_anzianita_cliente
Self
Mfa
RaccDirSMLa
Dist
Sport
−0.5
0.0
0.5
1.0

Gruppo 3


Gruppo 3 ($\overline{y}_{\text{gruppo 3}} = -21.715$)
ImpieghiSMLa
Eta
Mfa
RaccDirSMLa
Dist
RaccIndSMLa
Mesi_anzianita_cliente
Msa
Sport
Self
−0.5
0.0
0.5
1.0
1.5

Differenze tra i gruppi

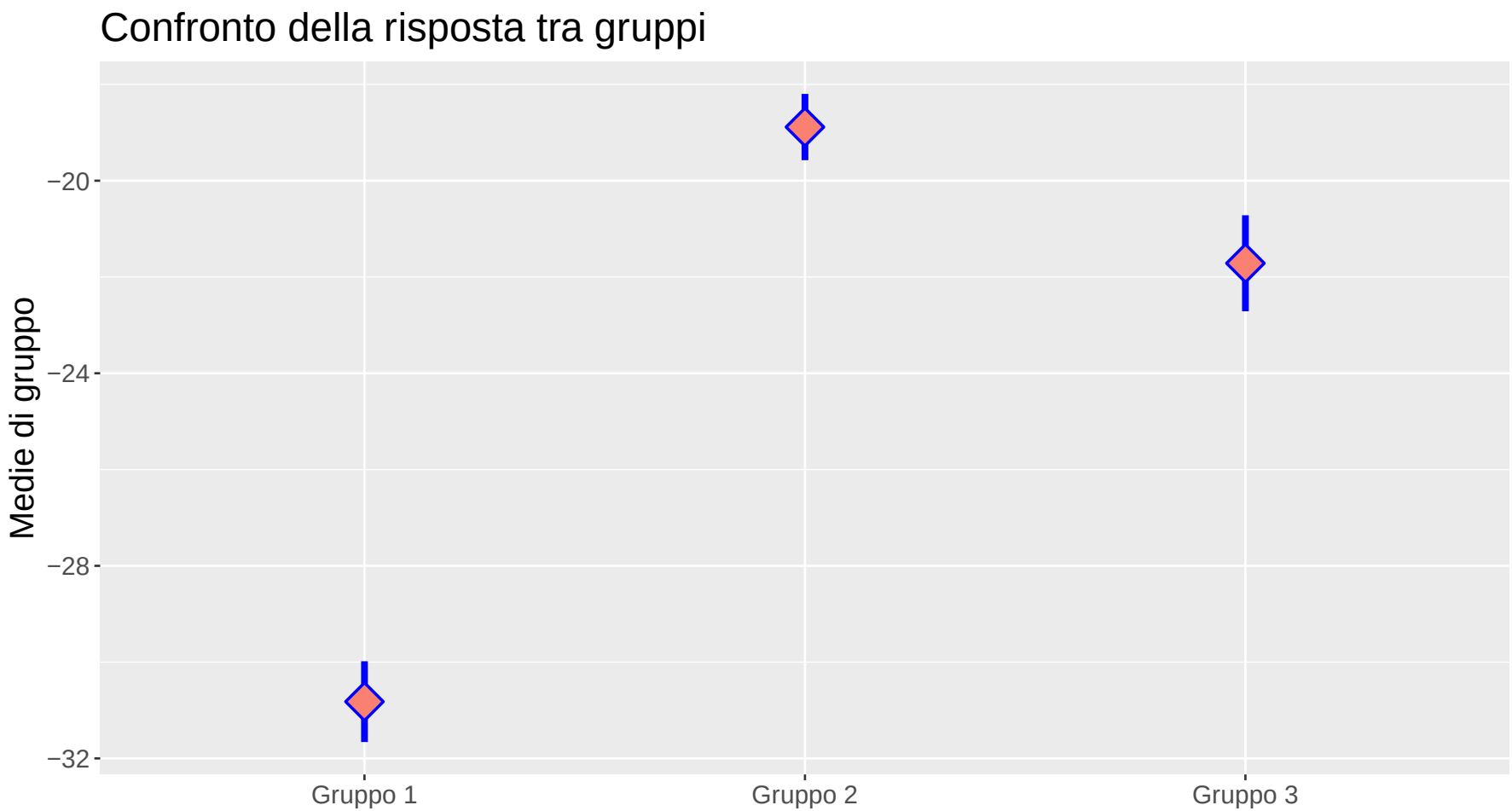


Figura: Confronto tra le medie di y tra i diversi gruppi

Commenti

I cluster sono così definiti:

- Gruppo 1: alto rischio → nonostante sia il gruppo più a rischio, la banca potrebbe non essere interessata a proporre nuove promozioni in quanto tale cluster presenta valori di `MFA` e `MSA` minori rispetto alla media.
- Gruppo 2: basso rischio → sono clienti che perdono pochi prodotti con un `MSA` elevato ma una `Età` avanzata, quindi la perdita del prodotto potrebbe essere una dinamica temporanea e poco preoccupante (anche a fini del Churn).
- Gruppo 3: medio rischio → vista l'`Età` e gli `Impieghi`, questo gruppo è caratterizzato da una clientela giovane e quindi da fidelizzare per il medio-lungo termine.

Sarà successivamente la banca a decidere quali promozioni proporre a ciascuno dei segmenti.

Conclusioni finali

Conclusioni finali

Lo studio condotto potrebbe servire alla prevenzione della perdita di prodotti da parte del cliente (scenario migliore), e al conseguente probabile abbandono dello stesso (scenario peggiore).
Sarebbe d'interesse effettuare l'analisi con un intervallo di tempo maggiore rispetto ad un anno, in modo tale da arginare il problema della presenza di zeri.
Inoltre, grazie alla segmentazione, risulta possibile proporre soluzioni *ad-hoc* per utenti simili.

Possibili sviluppi

A partire dall'informazione relativa alla numerosità dei prodotti posseduti da un cliente, sarebbe stato interessante disporre anche della tipologia di questi al fine di effettuare una *MBA - Market Basket Analysis*. Un ulteriore utilizzo della tipologia di prodotti posseduti da un cliente potrebbe esplicitarsi in un *collaborative filtering*:

1. individuazione di utenti simili in base alle covariate;
2. suggerimento di nuovi prodotti da acquisire in base alle similarità trovate al punto precedente.

GRAZIE PER L'ATTENZIONE