

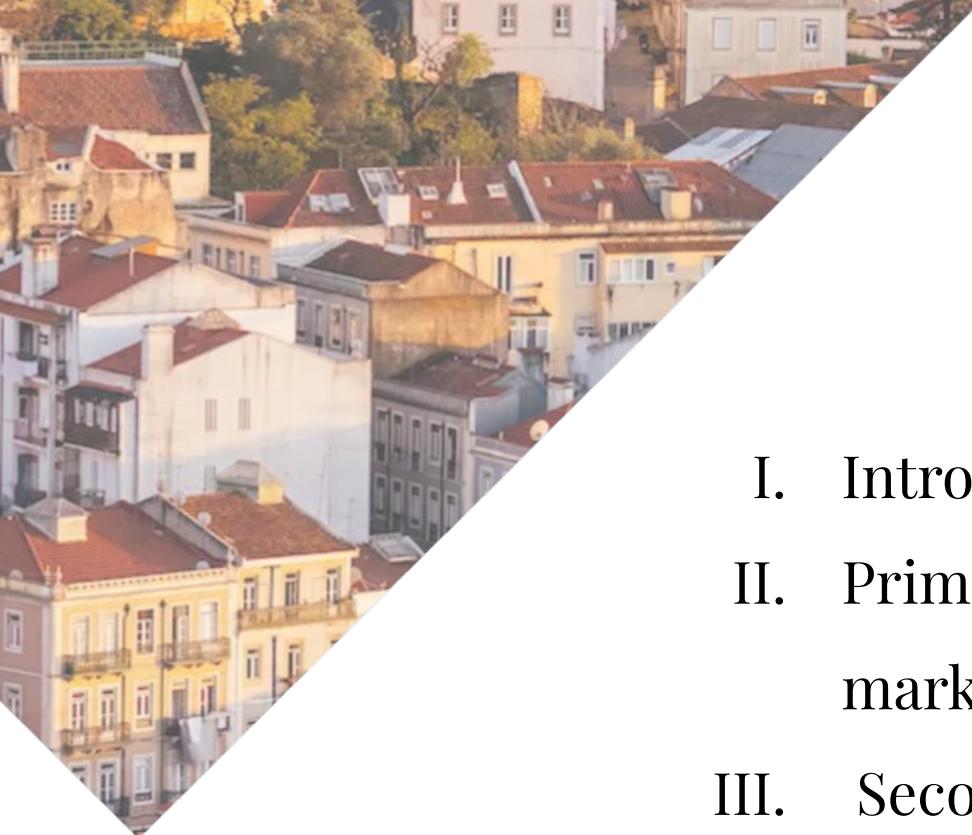


UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



# Bank Marketing

M. Pellizzari, A. Vignato, A. Ciandri



# Indice

- I. Introduzione al contesto e descrizione dei dati
- II. Prima domanda di business: analisi della campagna di marketing
- III. Seconda domanda di business: segmentazione della clientela
- IV. Terza domanda di business: analisi del fenomeno del default
- V. Conclusioni

# Contesto e obiettivi



Necessità del reparto marketing di un protocollo per affrontare eventuali crisi economiche o finanziarie future.

Studiare le campagne di marketing attuate in precedenti periodi di crisi .

Cercare di individuare dei meccanismi efficaci da sfruttare in periodi di crisi.

**L'ipotesi alla base del piano è che permanga una certa regolarità comportamentali dei clienti.**

# Descrizione dei dati



Campagna di  
telemarketing  
del 2008-2010

41118 chiamate  
21 variabili

Campagna di  
telemarketing  
precedente

45 221 chiamate  
17 variabili

Obiettivo:  
sottoscrizione  
deposito a m/l  
termine

# Descrizione dei dati

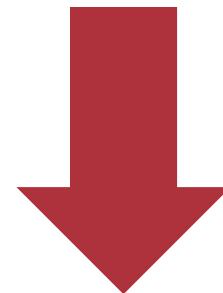
Variabili disponibili nel primo dataset

Caratteristiche del cliente	Informazioni sulla chiamata	Contatti precedenti	Attributi del contesto socio-economico
<ul style="list-style-type: none"><li>• Età</li><li>• Tipologia di lavoro</li><li>• Stato civile</li><li>• Livello di istruzione</li><li>• Status di default</li><li>• Mutuo sulla casa</li><li>• Prestito personale</li></ul>	<ul style="list-style-type: none"><li>• Tipo di contatto (telefono fisso o cellulare)</li><li>• Mese del contatto</li><li>• Giorno della settimana del contatto</li><li>• Durata della chiamata</li><li>• <b>Outcome della chiamata</b></li></ul>	<ul style="list-style-type: none"><li>• Nr. Contatti durante la campagna</li><li>• Nr. Contatti durante la campagna precedente</li><li>• Nr. Giorni passati dall'ultimo contatto della campagna precedente</li><li>• Outcome della campagna precedente</li></ul>	<ul style="list-style-type: none"><li>• Indice dei prezzi al consumo (mensile)</li><li>• Indice di fiducia dei consumatori (mensile)</li><li>• Euribor a 3 mesi (giornaliero)</li><li>• Numero di occupati (quadrimestrale)</li><li>• Variazione nel numero di occupati (quadrimestrale)</li></ul>



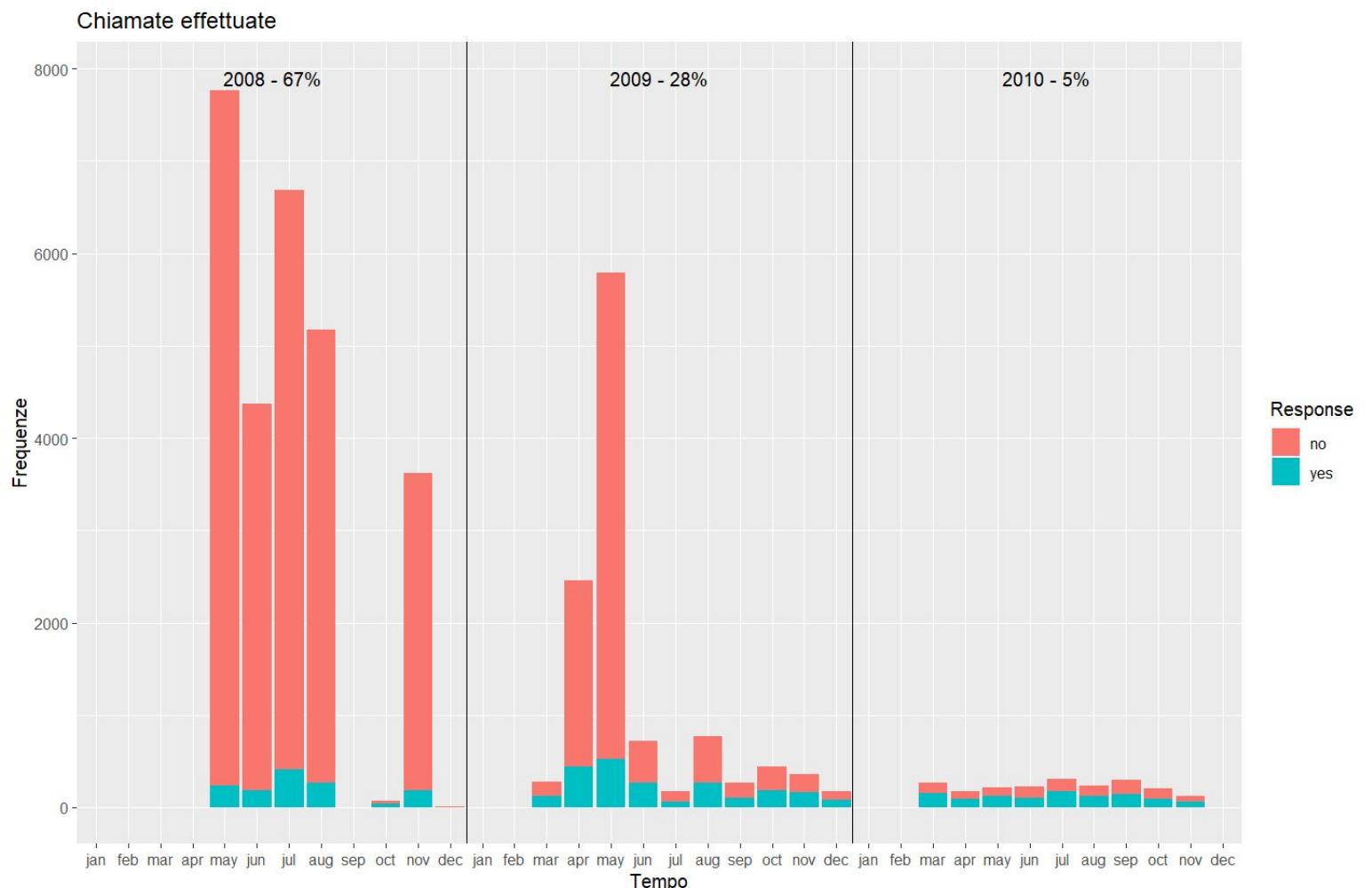
# Approccio al problema

Cercare di individuare le caratteristiche dei clienti che sono interessati al deposito e costruire un modello che preveda la probabilità di accettare l'offerta, tenendo conto anche del contesto socio-economico del Paese.

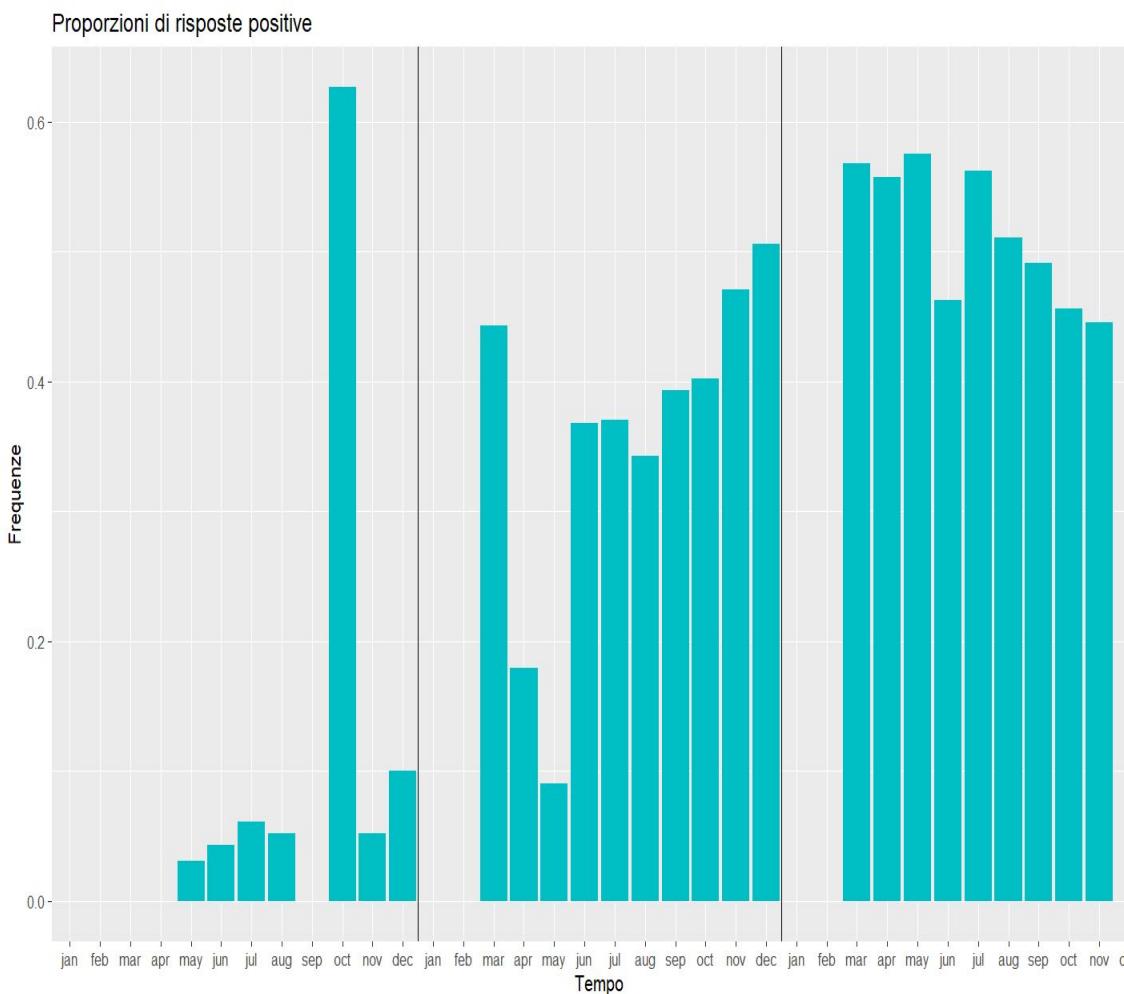


MODELLI DI CLASSIFICAZIONE

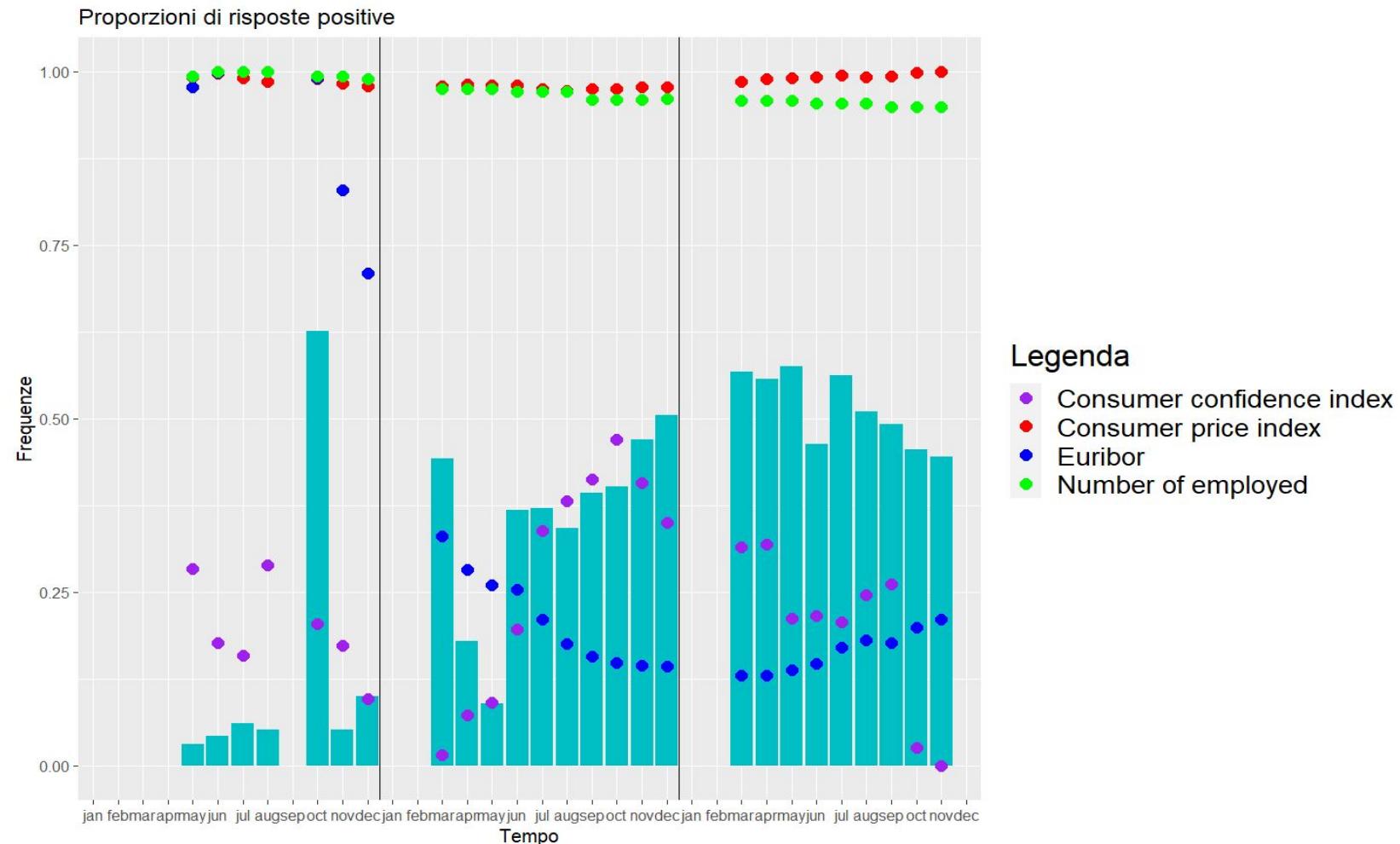
# Esplorazione del primo dataset



# Esplorazione del primo dataset

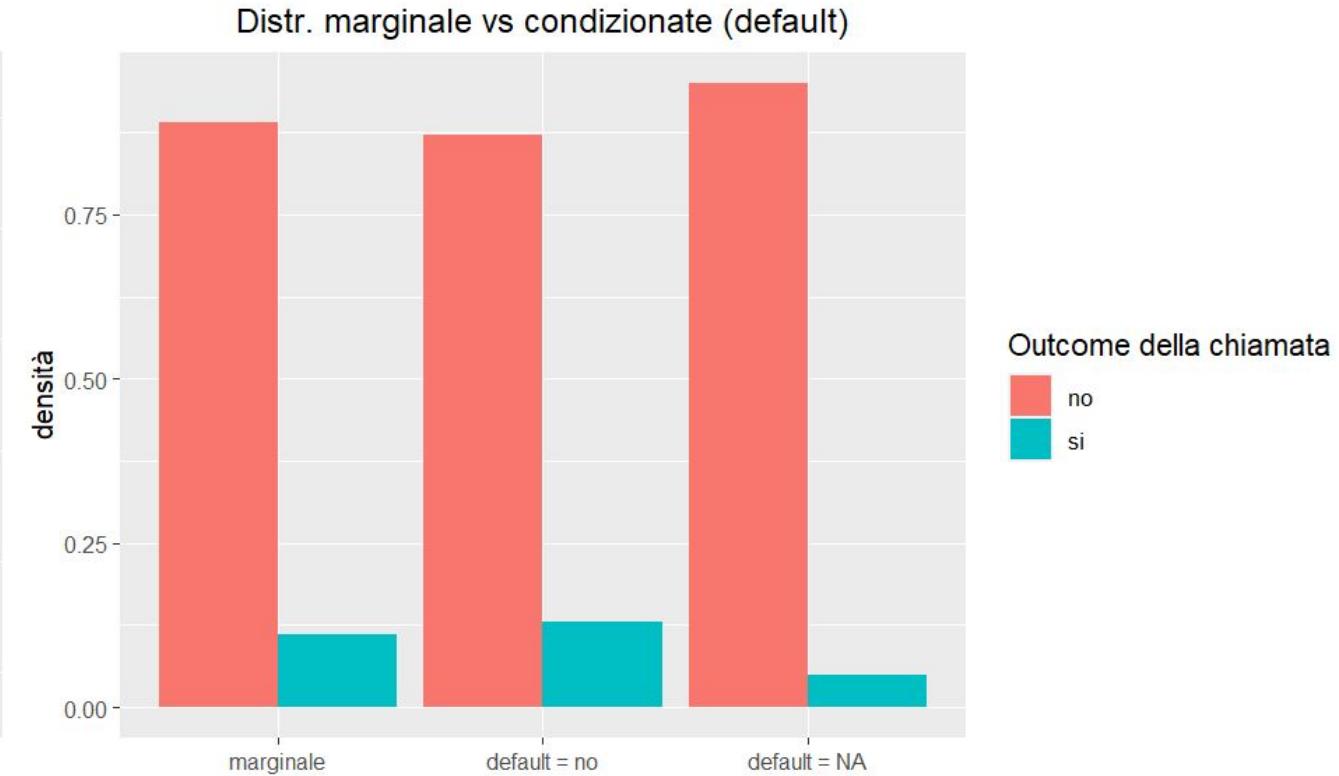
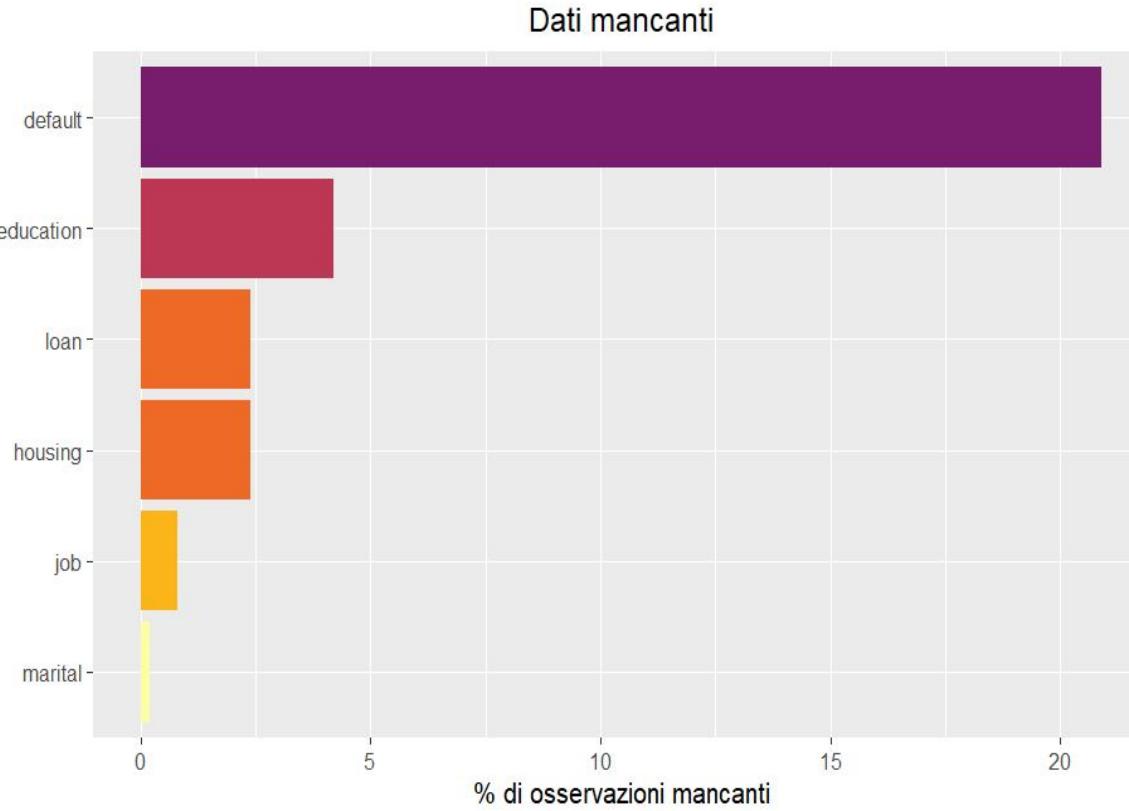


# Esplorazione del primo dataset



# Pulizia dei dati

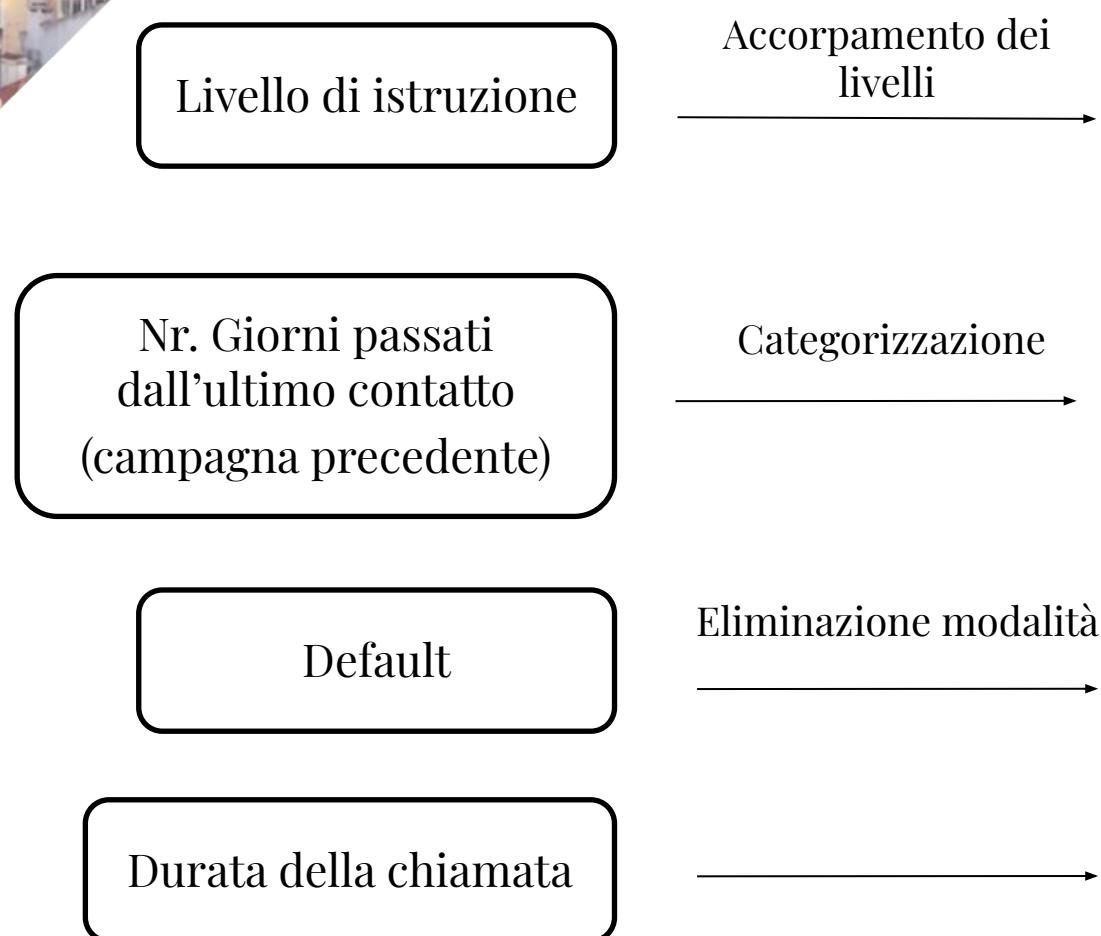
## Gestione dei dati mancanti





# Pulizia dei dati

## Trasformazioni delle variabili



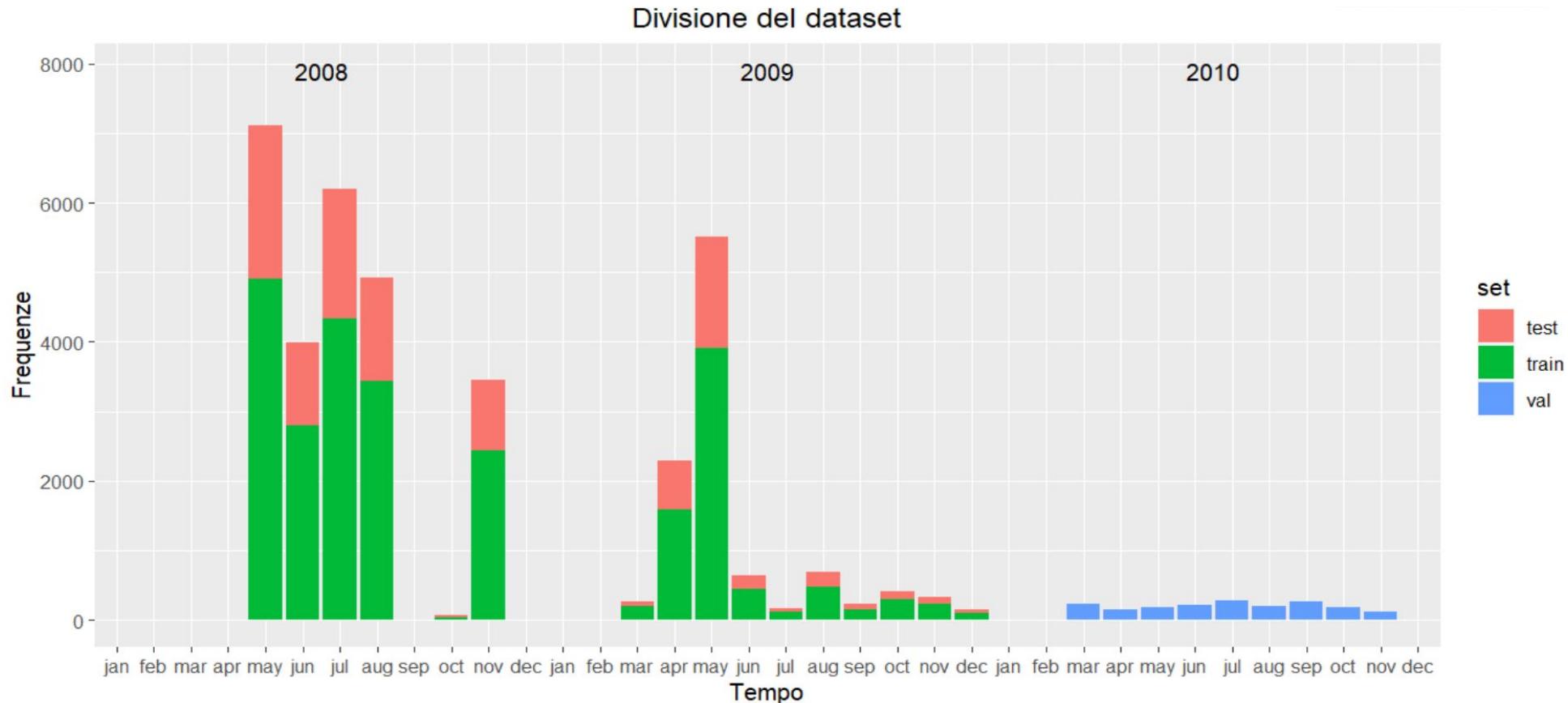
Accorpamento del livello «analfabeta» con il il livello di istruzione più bassa (3 anni di studio)

- Mai contattato
- Meno di 1 settimana
- Più di 1 settimana

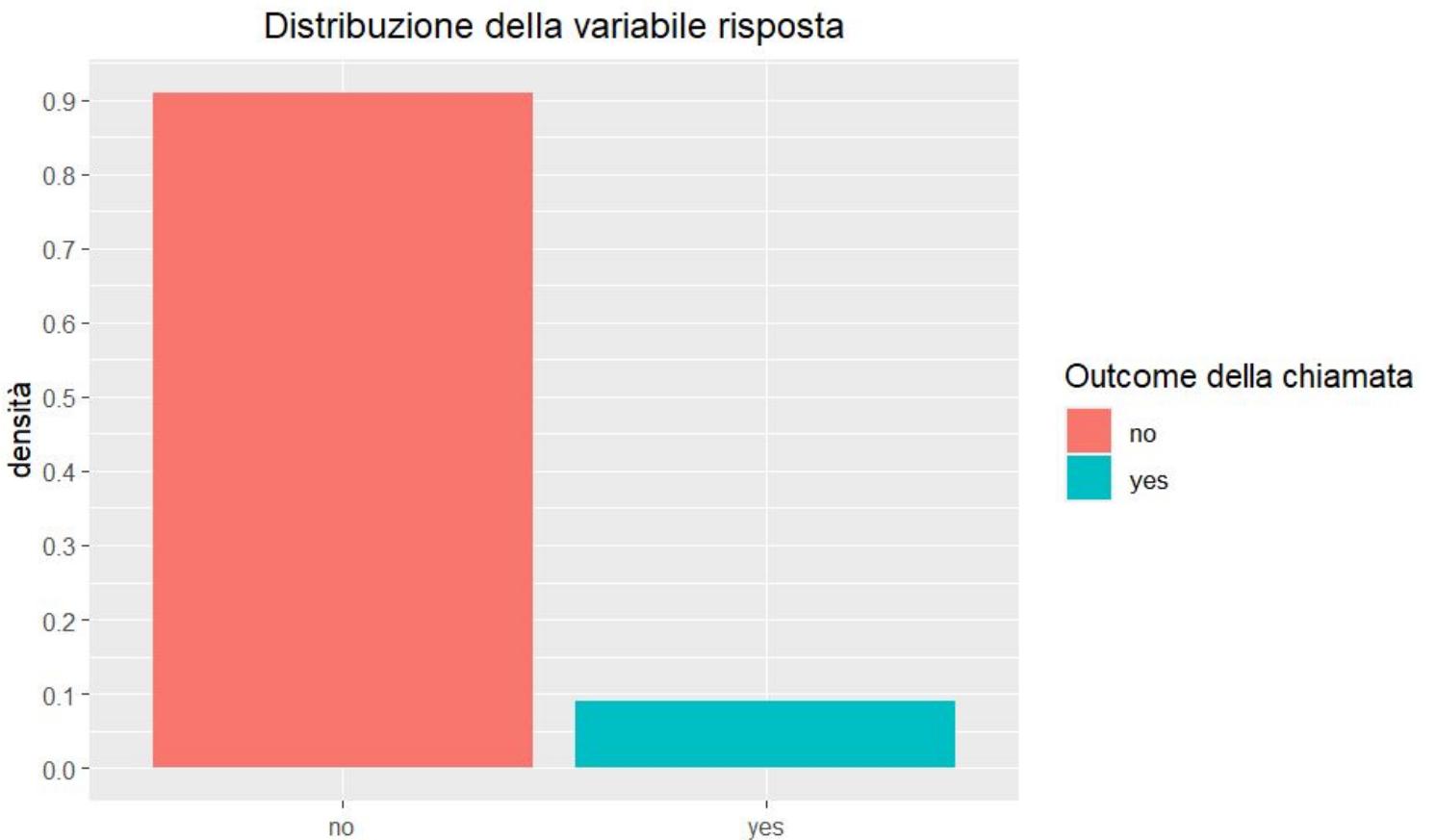
Eliminazione delle osservazioni con modalità “sì”, con sole 3 osservazioni.

Esclusa perché leaker

# Train set, test set e validation set



# Dataset sbilanciato



# È un problema?

Metriche modello logit (test set) - diversi metodi per bilanciare il dataset

Configurazione	Accuratezza	Precisione	Recupero	F1-score
Dataset sbilanciato	0.8872	0.3786	0.4417	0.4077
Pesi (3:1)	0.8851	0.3740	0.4562	0.4111
Sottocampionamento dei “no”	0.8881	0.3807	0.4354	0.4062
Sovracampionamento dei “sì”	0.8833	0.3674	0.4531	0.4058

# Scelta della soglia e metrica

Si è scelto di valutare le capacità previsive dei modelli in termini di **indice F1**. Per ogni modello è stato scelto il valore per la soglia che consentisse di massimizzare il suddetto indice facendo previsioni nel test set.

		Osservati	
		Rifiuta	Accetta
Previsti	Rifiuta	Da non contattare	Mancato guadagno
	Accetta	Spreco di risorse	Da contattare

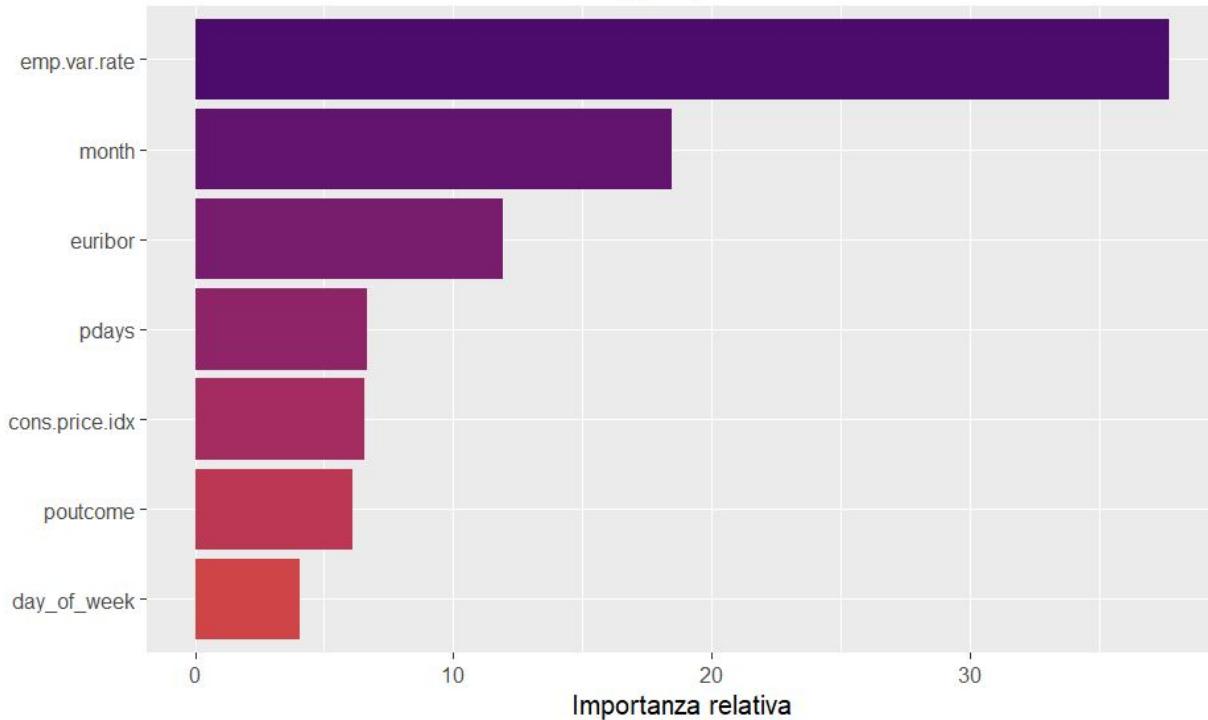
# Risultati

Modello	Accuracy	Precision	Recall	F1.score
Gradient Boosting	0.8882	0.3864	0.4625	<b>0.4211</b>
AdaBoost stumps	0.8876	0.3831	0.4573	<b>0.4169</b>
GAM	0.8906	0.3919	0.4438	<b>0.4162</b>
AdaBoost depth=2	0.8852	0.3756	0.4625	<b>0.4146</b>
Random Forest	0.8822	0.3672	0.4708	<b>0.4126</b>
PolyMARS	0.8875	0.3811	0.449	<b>0.4122</b>
Rete Neurale	0.8886	0.3839	0.4427	<b>0.4112</b>
Albero	0.8897	0.3865	0.4344	<b>0.409</b>
GLM	0.8872	0.3786	0.4417	<b>0.4077</b>
MARS additivo	0.8918	0.3918	0.4188	<b>0.4048</b>
LM con selezione variabili	0.8811	0.3612	0.4594	<b>0.4044</b>
LM completo	0.8867	0.3752	0.4354	<b>0.4031</b>
Bagging	0.8682	0.3113	0.4125	<b>0.3548</b>

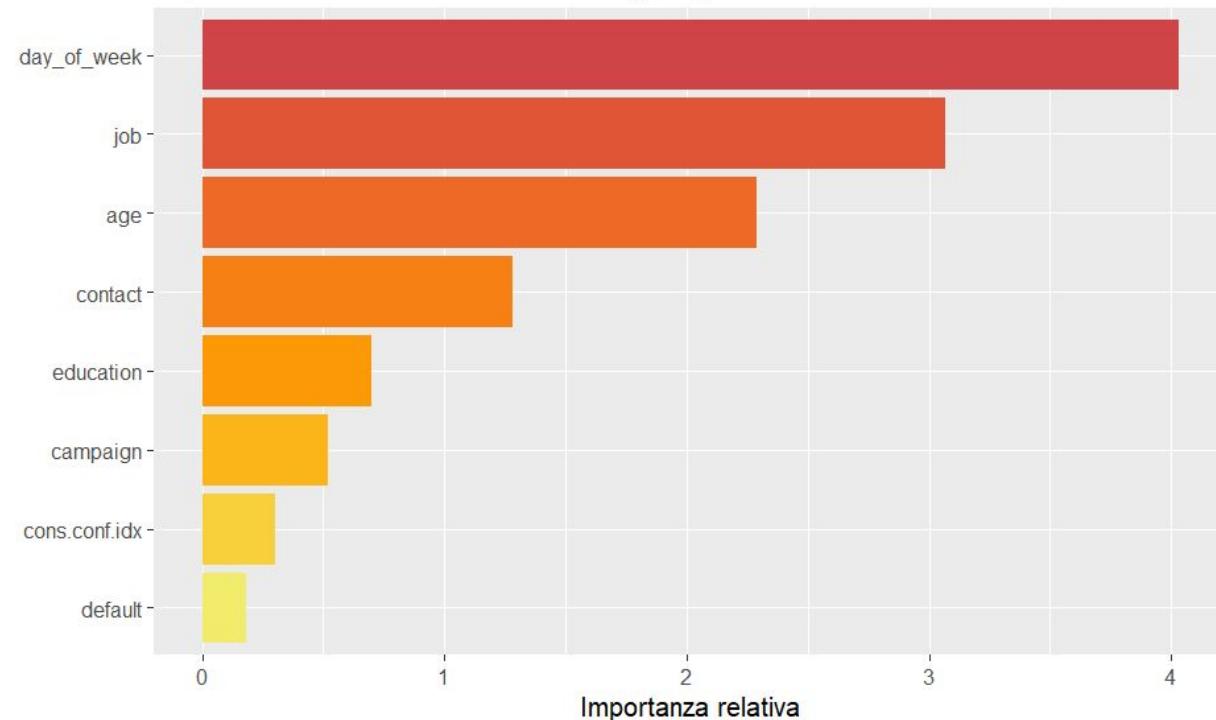


# Interpretazione dei risultati

Importanza relativa delle variabili (gbm)

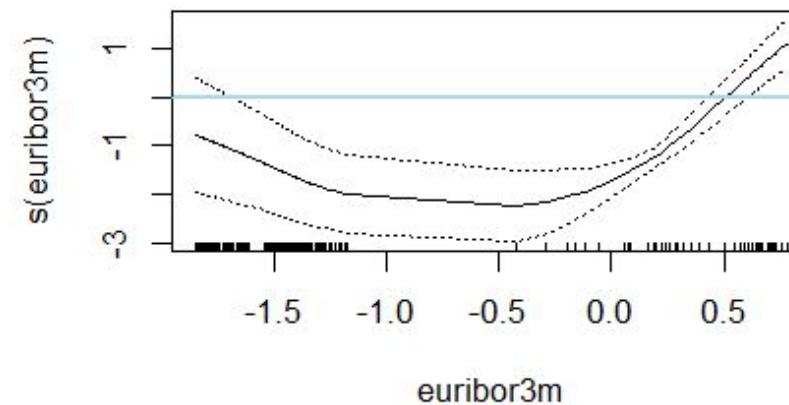
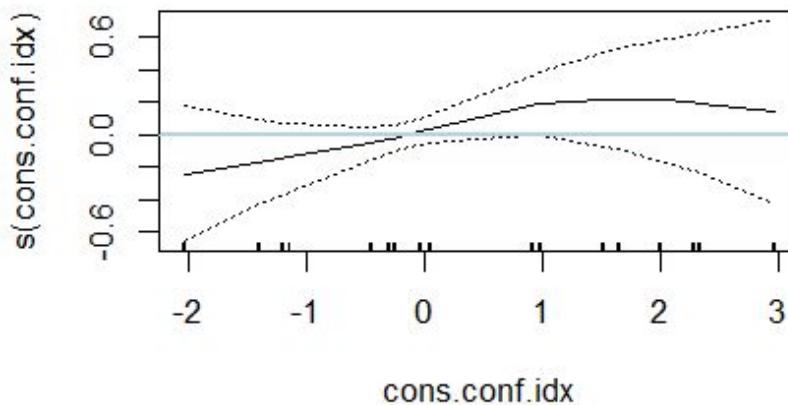
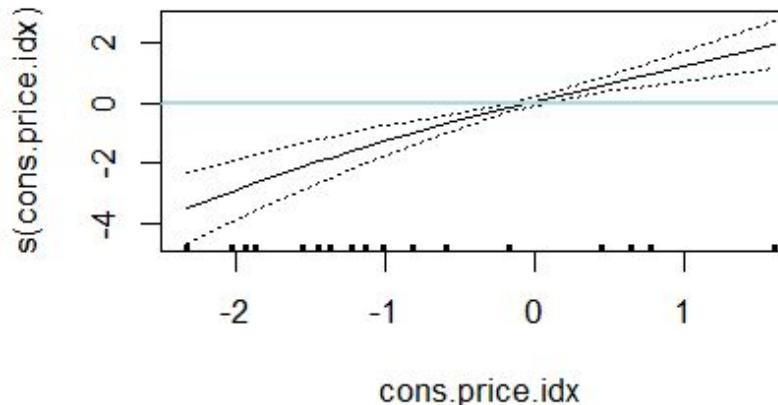
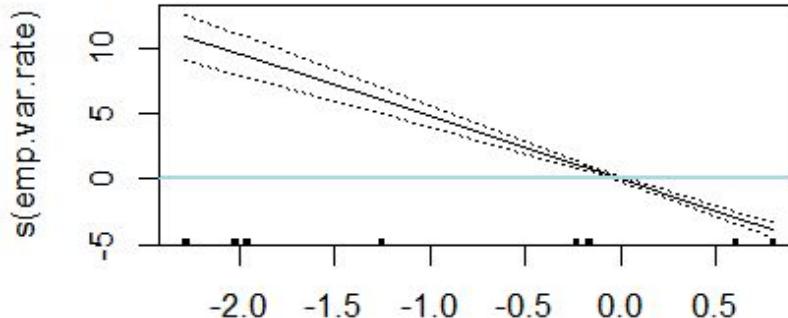


Importanza relativa delle variabili (gbm)

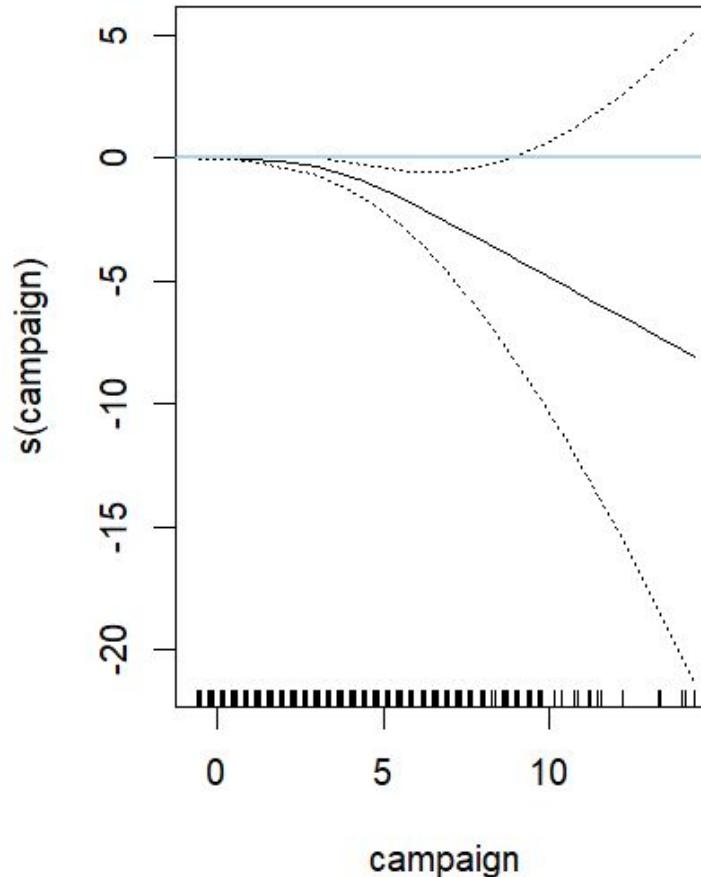
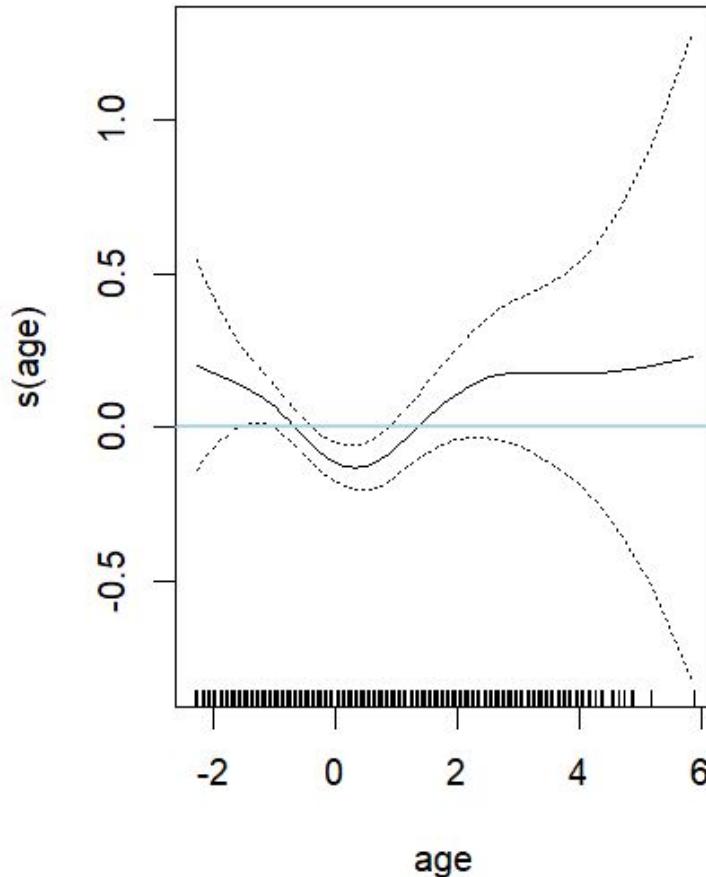


Si può osservare che le caratteristiche personali del cliente hanno meno importanza rispetto alle altre variabili del dataset.

# Effetti delle variabili socio-economiche

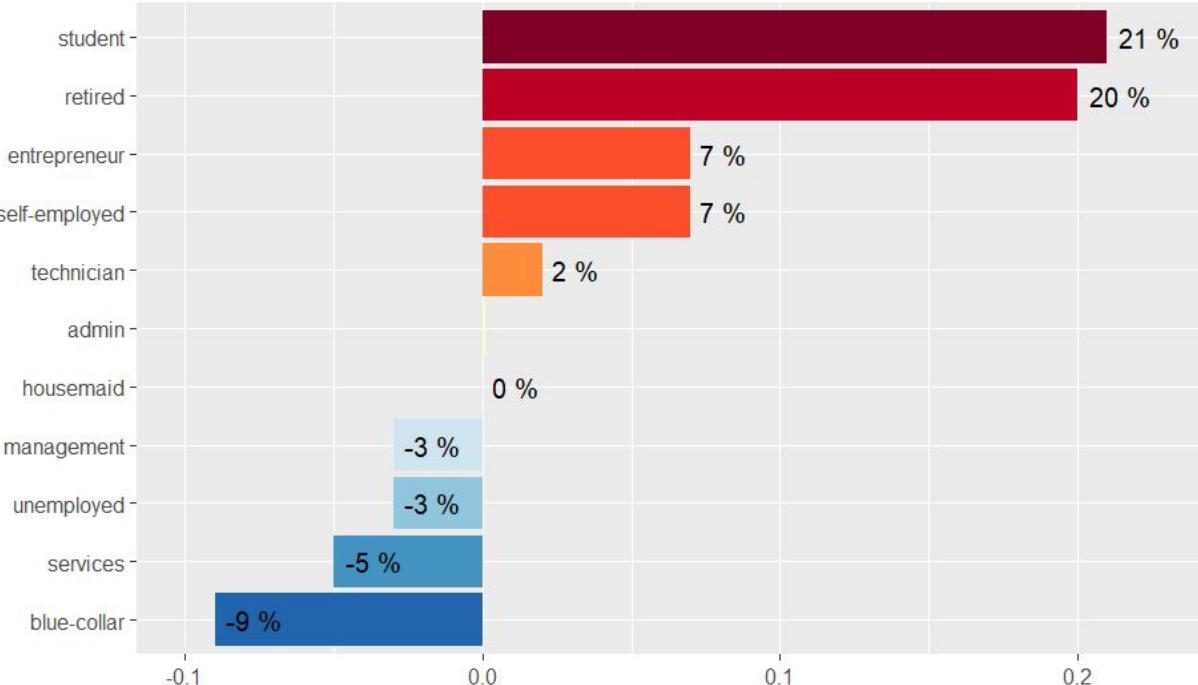


# Effetti dell'età e del numero di contatti durante la campagna corrente

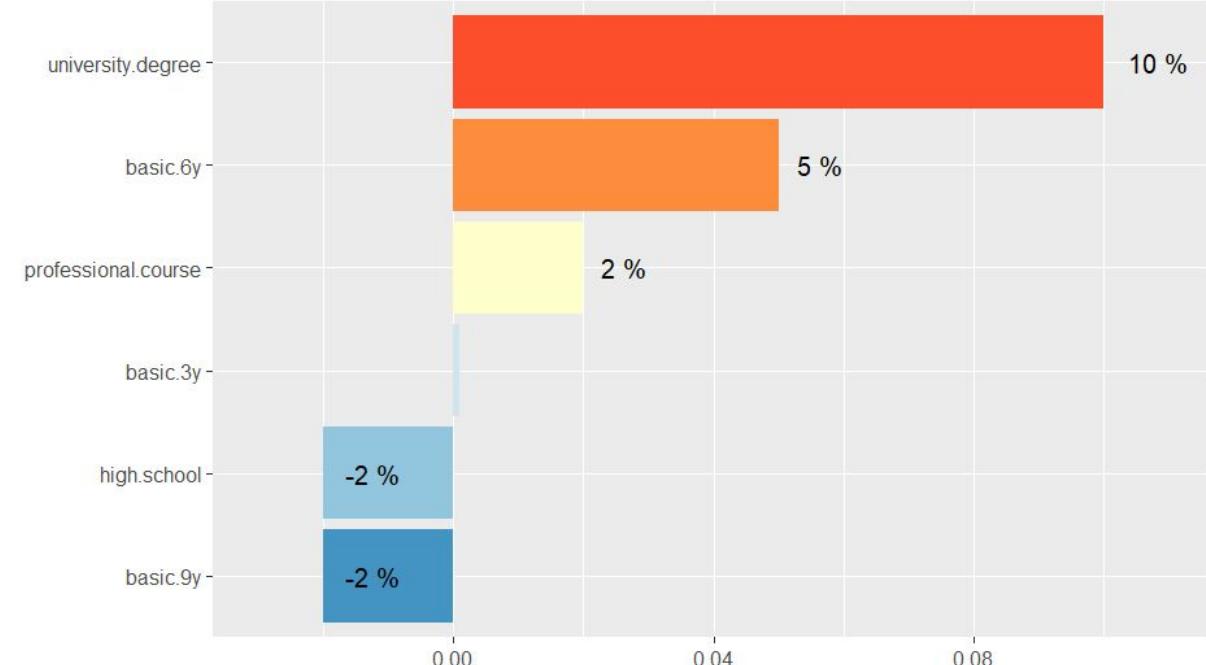


# Effetti del lavoro e dell'istruzione

Effetti del lavoro sulla quota

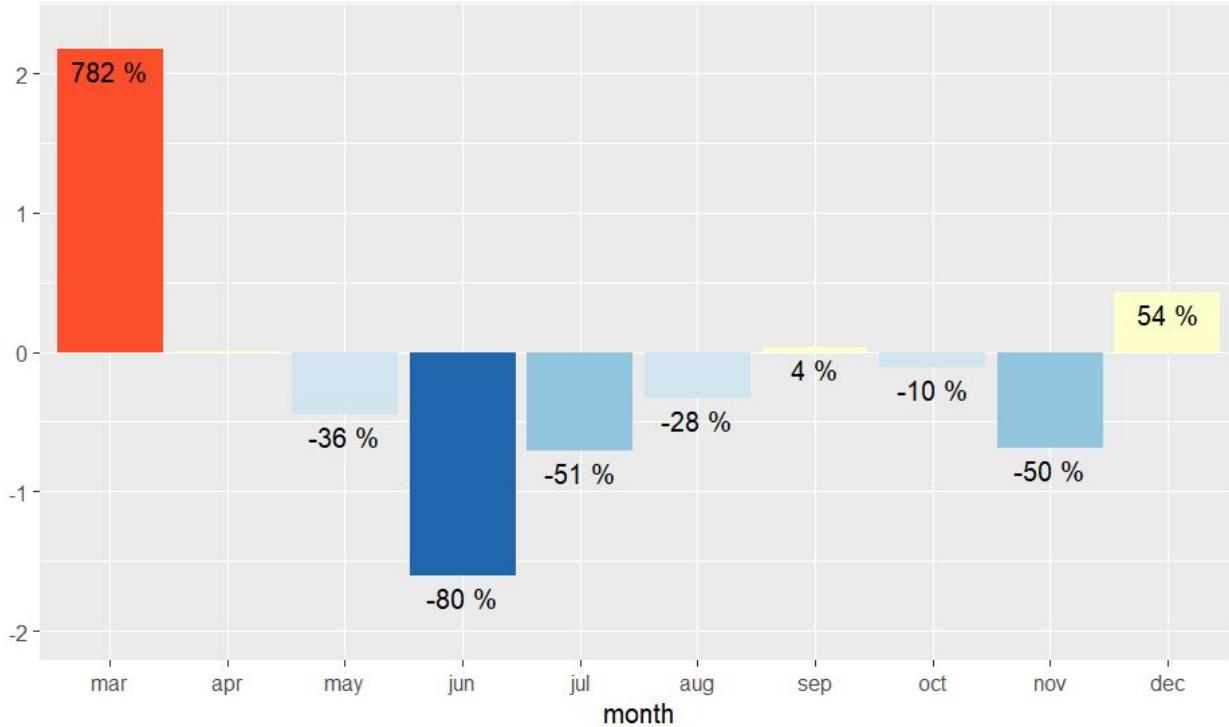


Effetti dell'istruzione sulla quota

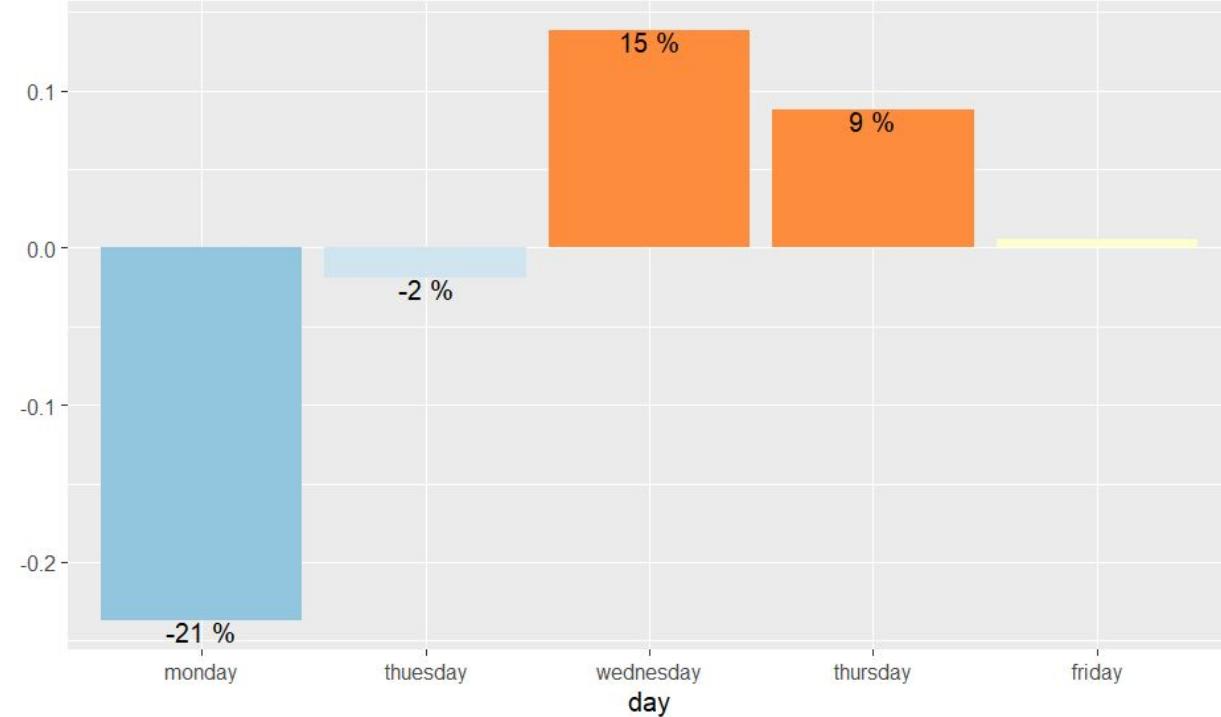


# Effetti del mese e del giorno della settimana

Effetti del mese sulla log-quota



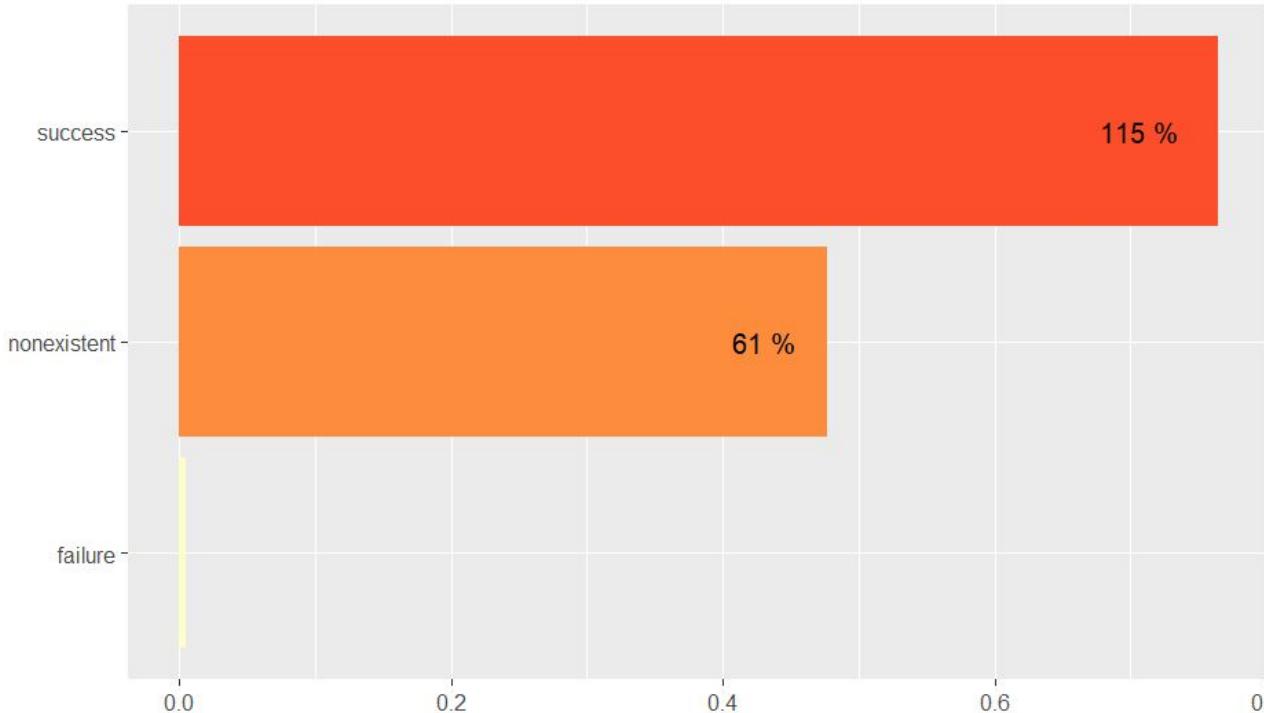
Effetti del giorno sulla log-quota



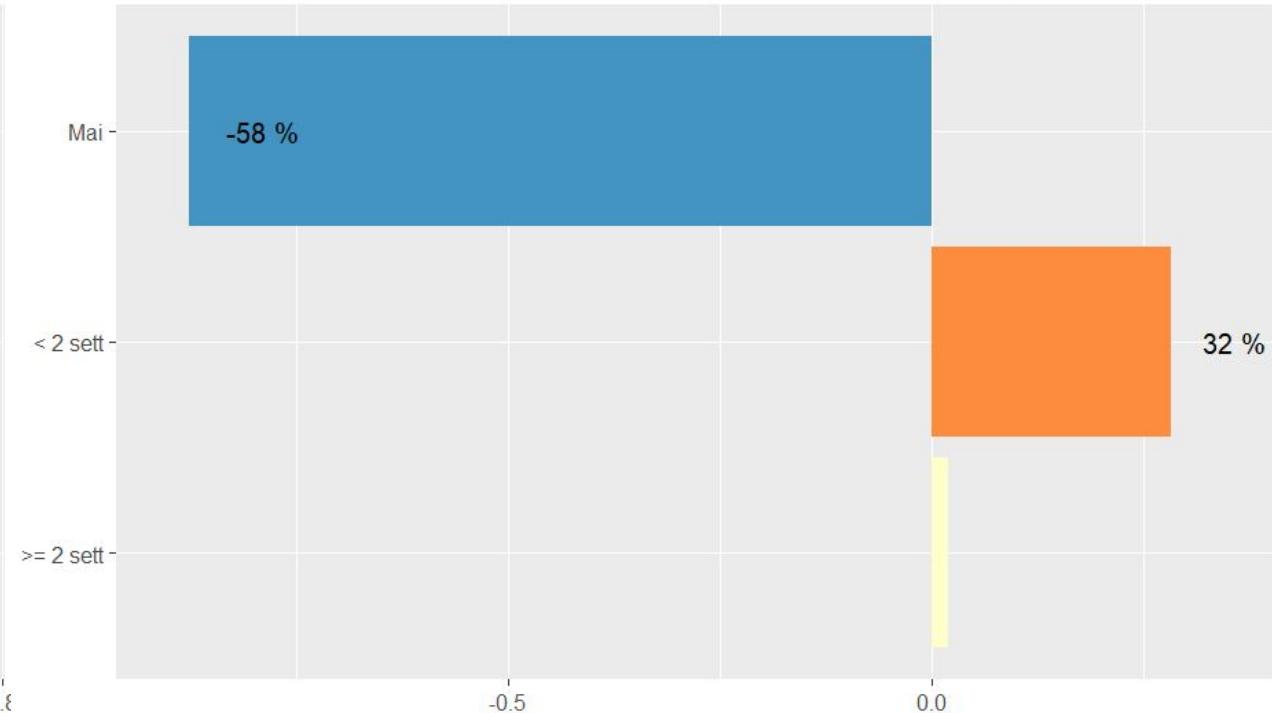
Al fine di rendere questi grafici più leggibili, gli effetti rappresentati sono i coefficienti stimati dal modello.

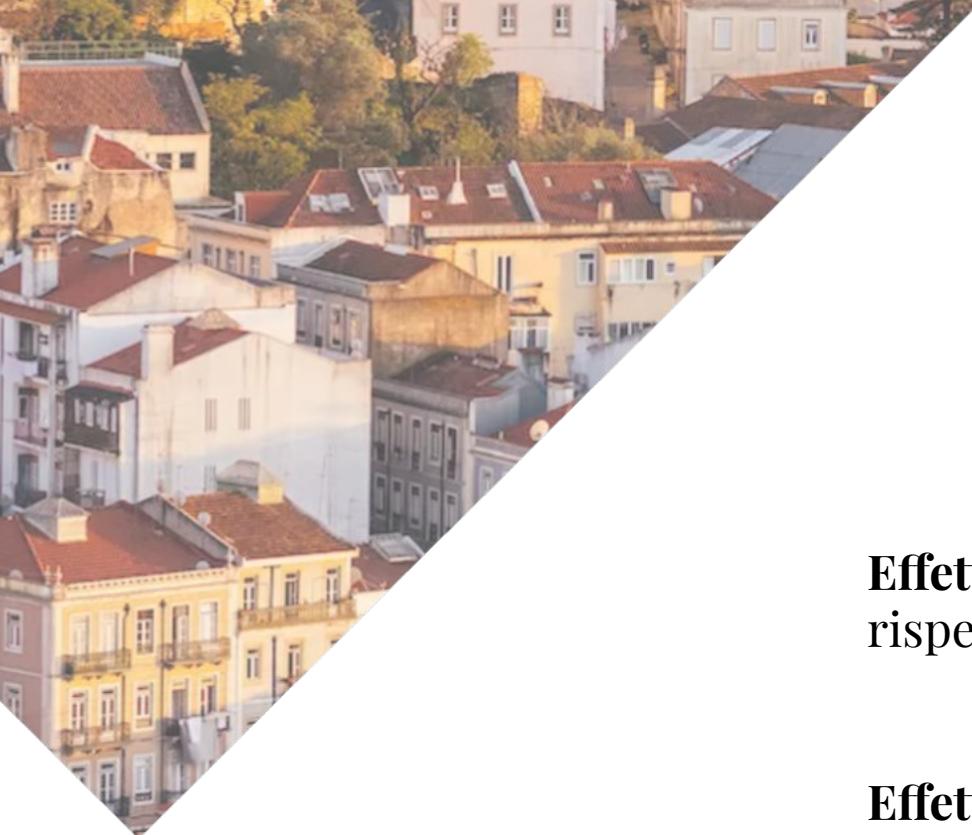
# Effetti della campagna precedente

Effetti della campagna precedente sulla risposta



Effetti del n° giorni trascorsi dall'ultimo contatto per la campagna precedente



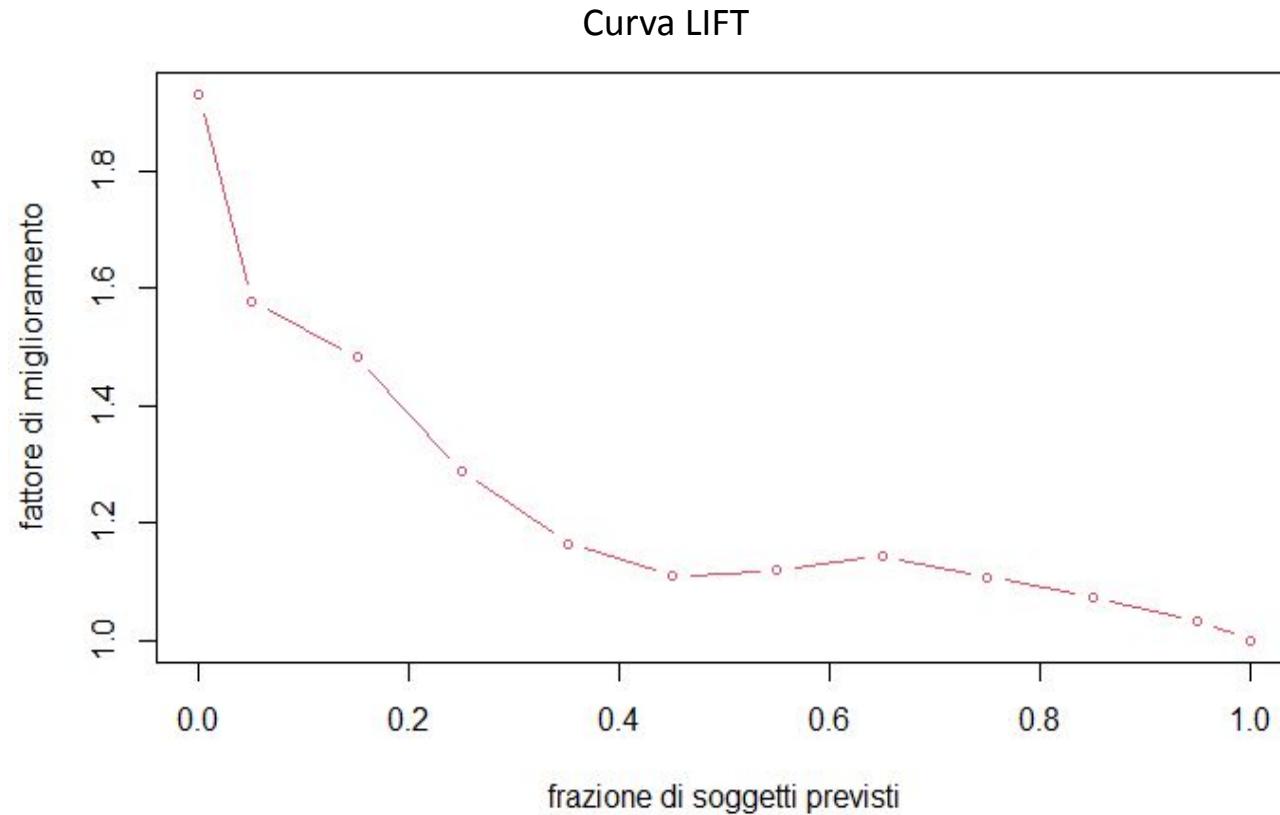


# Rimanenti variabili

**Effetto del default:** Non dichiarare il default riduce la quota del 21% rispetto al non essere in default.

**Effetto del tipo di contatto:** Contattare il cliente tramite cellulare riduce la quota del 42%, rispetto ad una chiamata sul telefono fisso.

# Risultati sul validation set - GBM

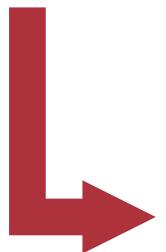


	Accuracy	Precision	Recall	F1-score
Validation set	0.6017	0.5835	0.8095	0.6781



# Segmentazione della clientela

Al fine di fornire un quadro più esauriente sul comportamento dei clienti in periodo di crisi, abbiamo completato lo studio con un'**analisi di segmentazione**. Vorremmo in questo modo individuare i gruppi di clienti più interessati al prodotto bancario offerto e le loro caratteristiche.



Per farlo, abbiamo utilizzato solo un sottoinsieme ridotto delle variabili a disposizione

# Segmentazione della clientela

Variabili utilizzate per la segmentazione

Caratteristiche del cliente	Informazioni sulla chiamata	Contatti precedenti	Attributi del contesto socio-economico
<ul style="list-style-type: none"><li>• Età</li><li>• Tipologia di lavoro</li><li>• Stato civile</li><li>• Livello di istruzione</li><li>• Status di default</li><li>• Mutuo sulla casa</li><li>• Prestito personale</li></ul>	<ul style="list-style-type: none"><li>• Tipo di contatto (telefono fisso o cellulare)</li><li>• Mese del contatto</li><li>• Giorno della settimana del contatto</li><li>• Durata della chiamata</li><li>• <b>Outcome della chiamata</b></li></ul>	<ul style="list-style-type: none"><li>• Nr. Contatti durante la campagna</li><li>• Nr. Contatti durante la campagna precedente</li><li>• Nr. Giorni passati dall'ultimo contatto della campagna precedente</li><li>• <b>Outcome della campagna precedente</b></li></ul>	<ul style="list-style-type: none"><li>• Indice dei prezzi al consumo (mensile)</li><li>• Indice di fiducia dei consumatori (mensile)</li><li>• Euribor a 3 mesi (giornaliero)</li><li>• Numero di occupati (quadrimestrale)</li><li>• Variazione nel numero di occupati (quadrimestrale)</li></ul>



# Trasformazione delle variabili

Riduzione dei livelli

Livello di istruzione  
(7 Livelli)

Accorpamento

- Istruzione bassa
- Istruzione media
- Istruzione alta

Tipologia di lavoro  
(11 Livelli)

Accorpamento

- Lavoro dirigenziale
- Lavoro manuale
- Lavoro impiegatizio
- Servizi
- Non occupato

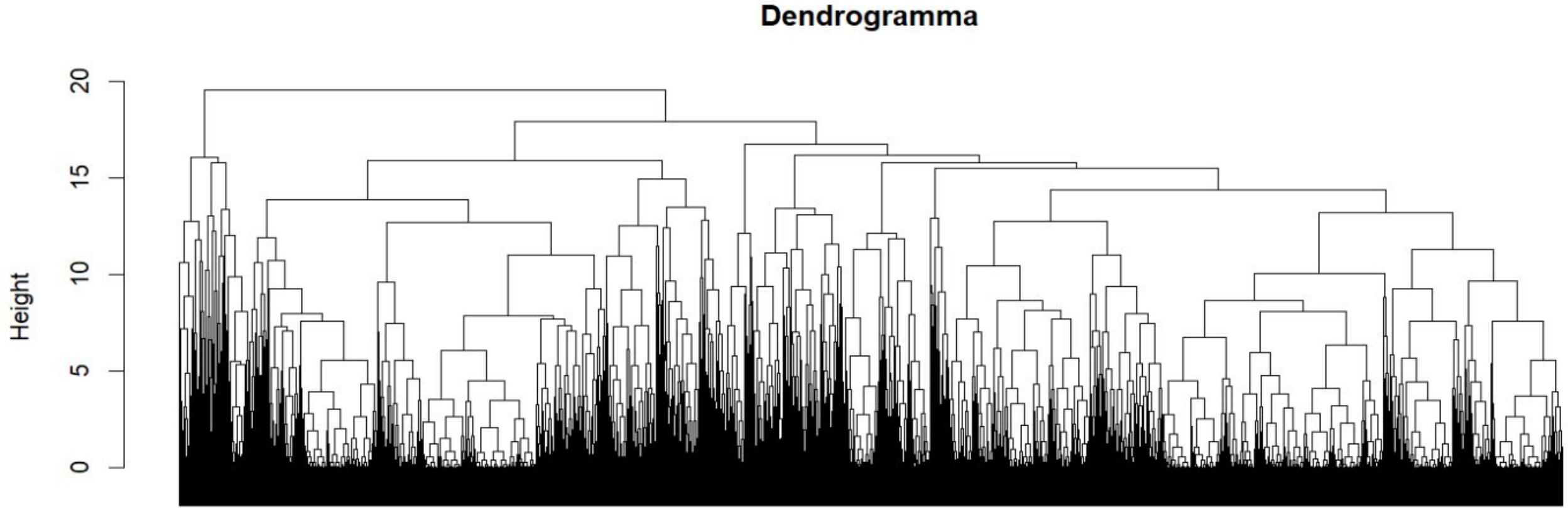


# Clustering gerarchico

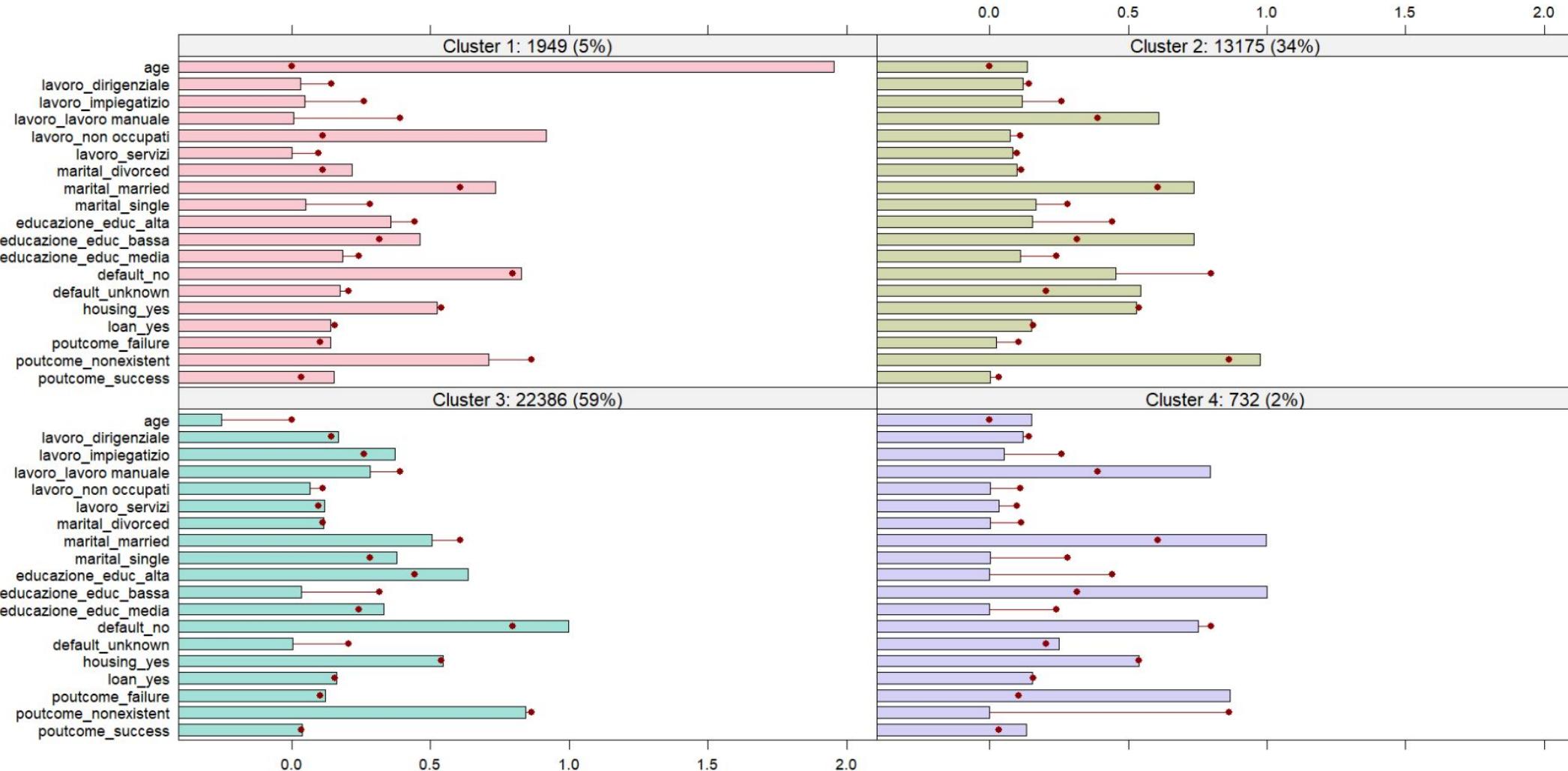
È quindi stata costruita la matrice di distanze usando la **distanza di Manhattan**. Per il clustering è stato utilizzato il metodo del **legame completo**.

Per costruire la matrice di distanze necessaria, ciascuna delle variabili qualitative disponibili è stata trasformata in un insieme di **variabili dummy**, che sono state **trattate come variabili quantitative**. La variabile “Età” è stata inoltre standardizzata.

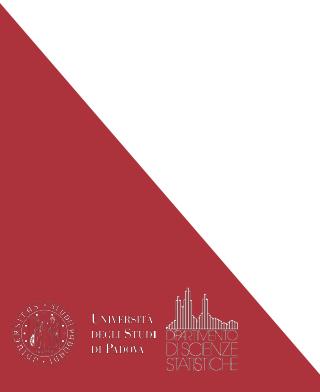
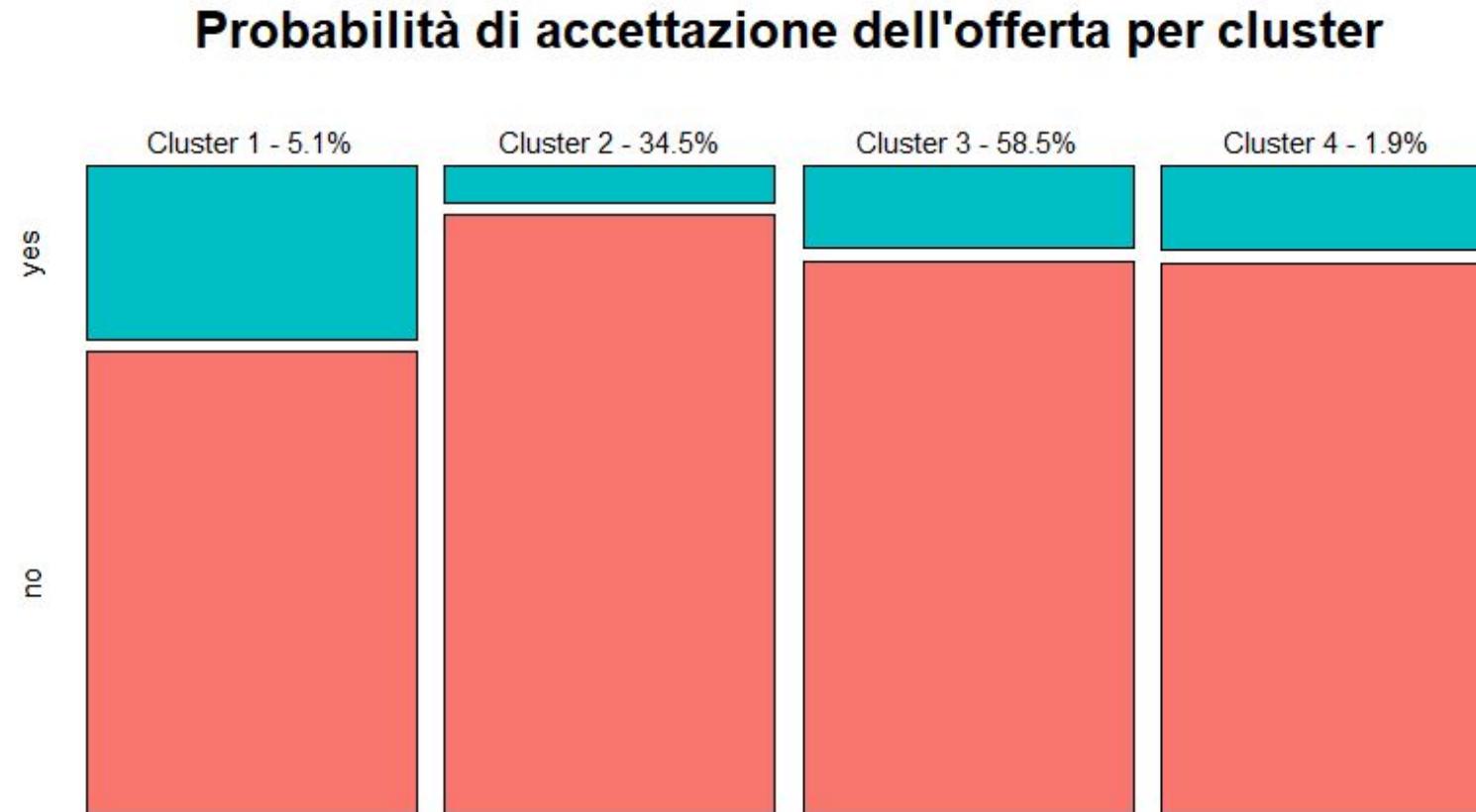
# Risultati - Dendrogramma



# Risultati - Caratteristiche dei cluster



# Risultati - Comportamento dei cluster





# Analisi del default

Un ultimo aspetto dell'analisi svolta riguarda il primo problema studiato. Nel prevedere la risposta del cliente all'offerta risultava infatti rilevante la variabile relativa allo **status di default**.

La variabile però aveva solo, di fatto, 2 modalità: “no default” e “**status ignoto**”, e la seconda risultava **ridurre significativamente** la probabilità di accettare l'offerta.

L'obiettivo della terza analisi è quindi quello di capire se gli individui che non dichiarano lo status di default sono in default, e cercare di capire chi questi sono.

Per farlo è stato usato il secondo dataset.



# Variabili disponibili

Il secondo dataset, che fa riferimento ad una campagna di marketing precedente, è molto simile al primo in termini di contenuti. Le variabili utilizzate (un sottoinsieme) sono le seguenti.

## Caratteristiche del cliente

- Età
- Tipologia di lavoro
- Stato civile
- Livello di istruzione
- **Status di default**
- Mutuo sulla casa
- Prestito personale
- Saldo del conto corrente

Le differenze principali con il dataset precedente sono che lo status di default **non ha valori mancanti** e che è disponibile il saldo del conto corrente.

# Preparazione dei dati

I dati mancanti presentano un pattern simile a quello del primo dataset, e per questo **le osservazioni relative sono state rimosse**. Sono state rimosse circa 2000 osservazioni.

Saldo del conto corrente

Costruzione di una nuova variabile

Saldo del conto corrente (**viene mantenuta**)



Dummy:

- Saldo negativo
- Saldo a zero
- Saldo positivo



# Approccio

L'obiettivo è quello di costruire dei modelli che permettano di capire quali sono gli individui che si trovano in default.

Dataset (43193 osservazioni)

Suddiviso in

**Training set**  
(60%)

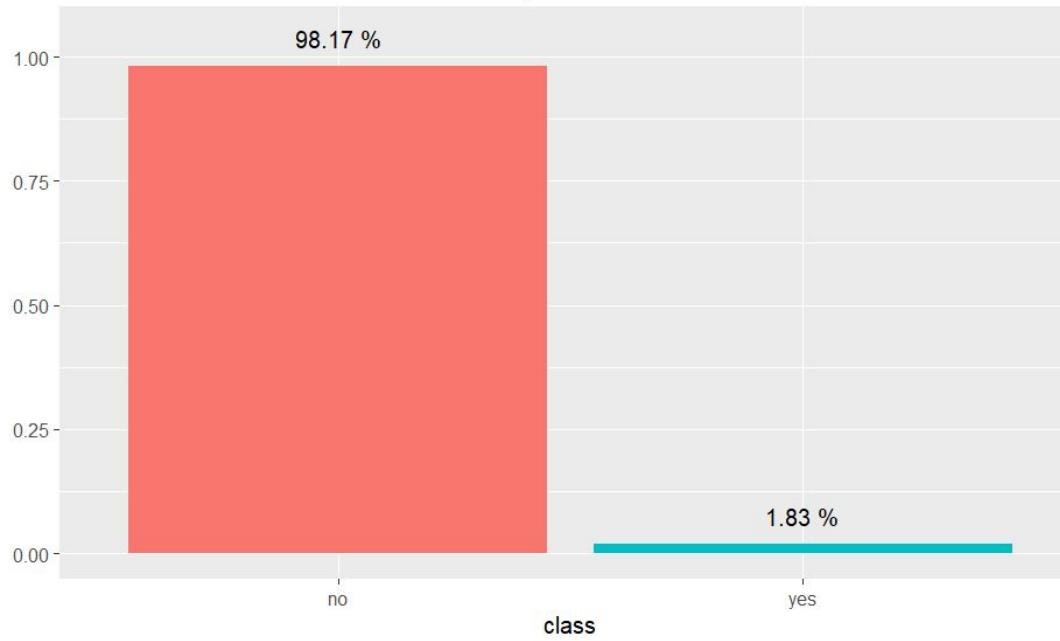
**Test set**  
(20%)

**Validation set**  
(20%)

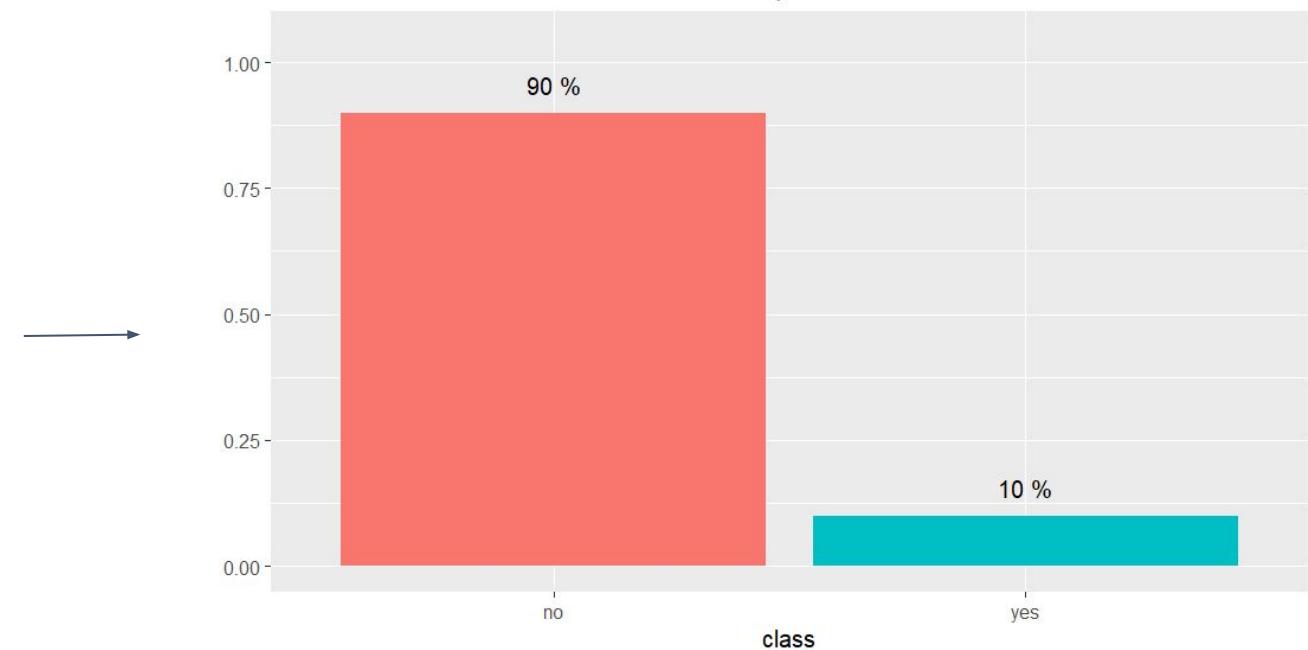
# Bilanciamento della risposta

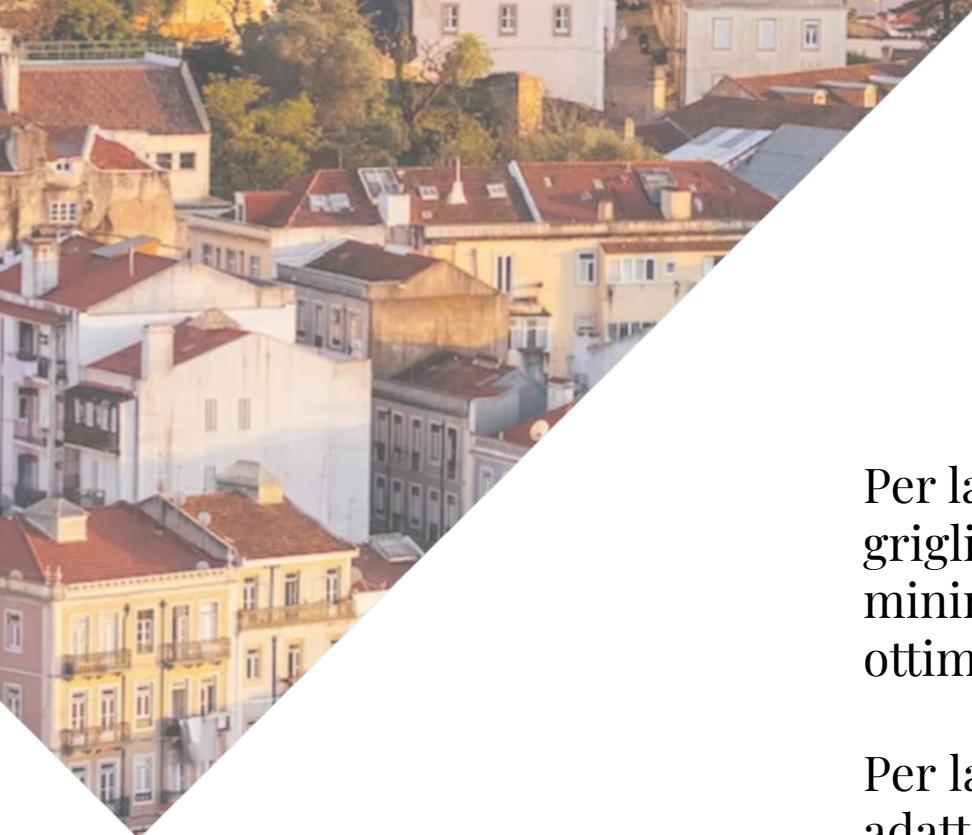
Anche in questo caso, la variabile risposta non era bilanciata, come si vede dal grafico sottostante. La variabile è stata quindi bilanciata (nel training set) tramite la combinazione di **sovracampionamento** (dei “si”) e **sottocampionamento** (dei “no”). Il training set ottenuto contiene 11000 osservazioni.

Distribuzione della variabile "default" prima



Distribuzione della variabile "default" dopo





# Soglia e metrica

Per la scelta della soglia da utilizzare è stato utilizzato il **test set**: data una griglia di possibili soglie, per ciascun modello è stata scelta quella che minimizza la distanza con **il punto (0,1)** nella **curva ROC** (il punto ottimale).

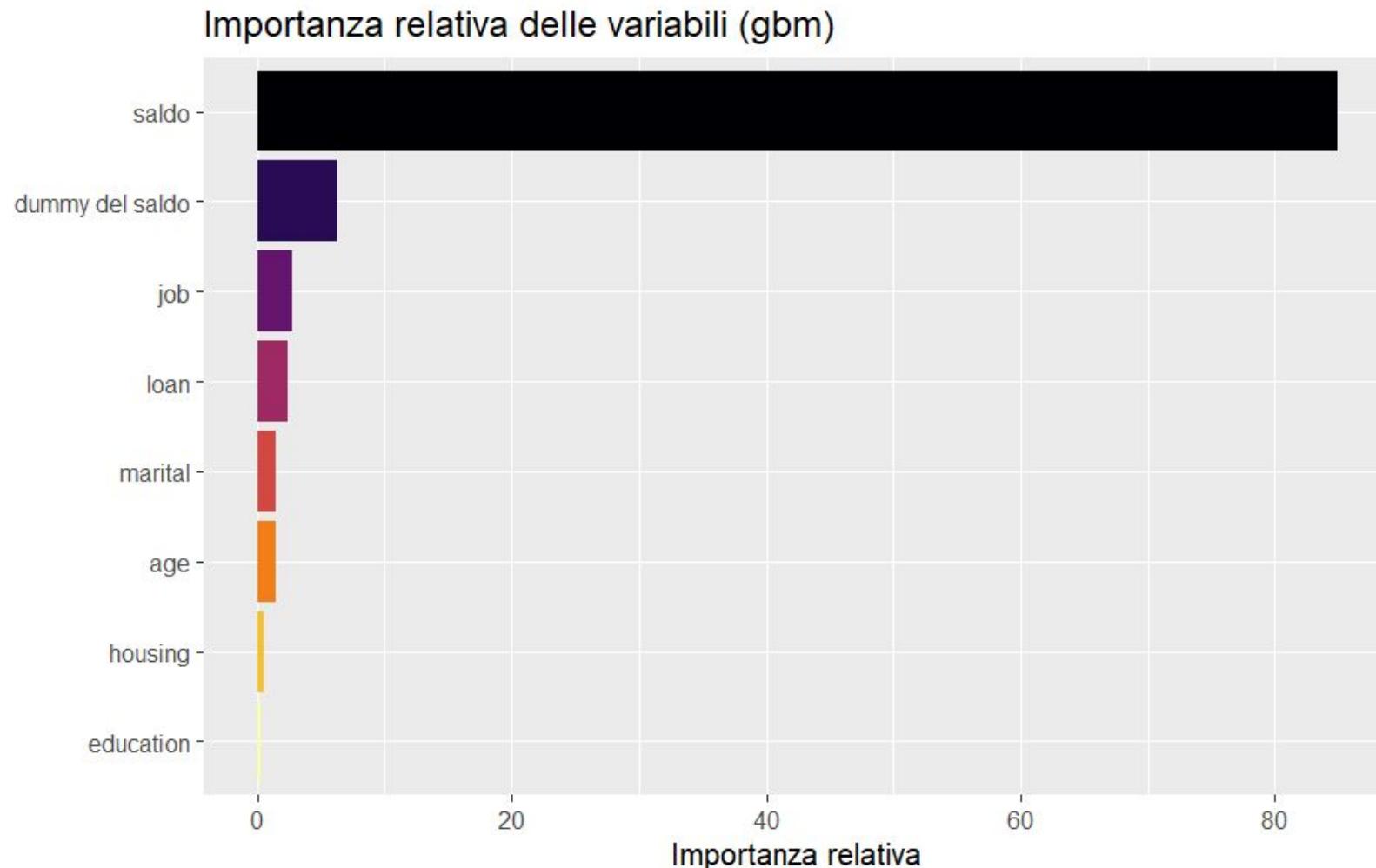
Per la scelta del modello migliore è stato invece utilizzato l'**indice F1**, più adatto all'accuratezza visto lo sbilanciamento presente nel dataset. Quest'ultimo è stato calcolato, per ciascun modello, sul **validation set**, utilizzando la soglia individuata nel test set.

# Risultati

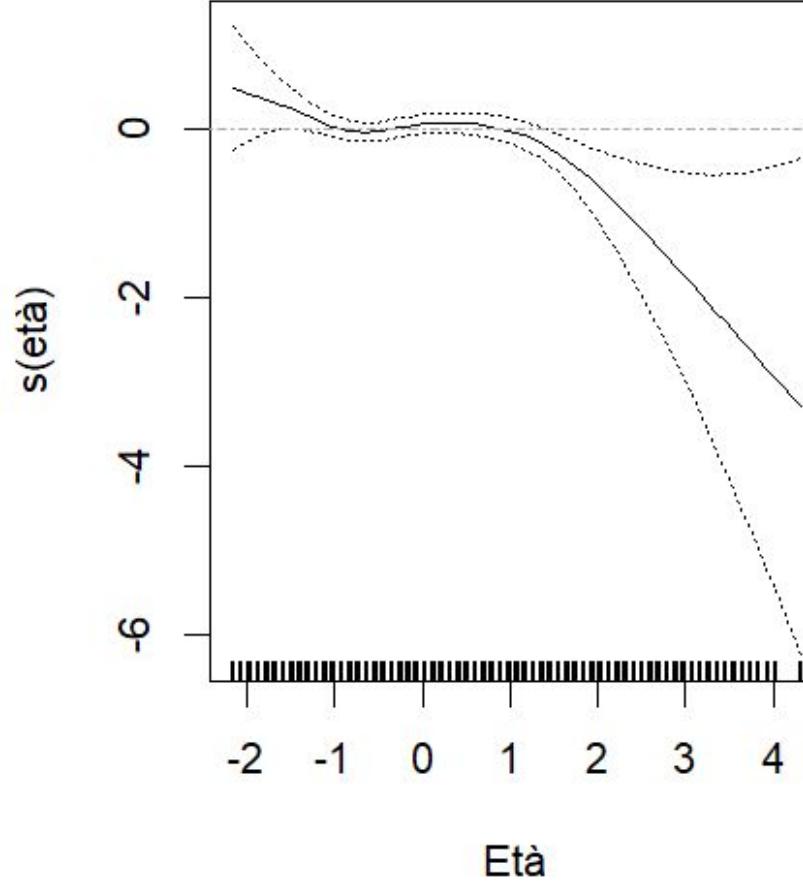
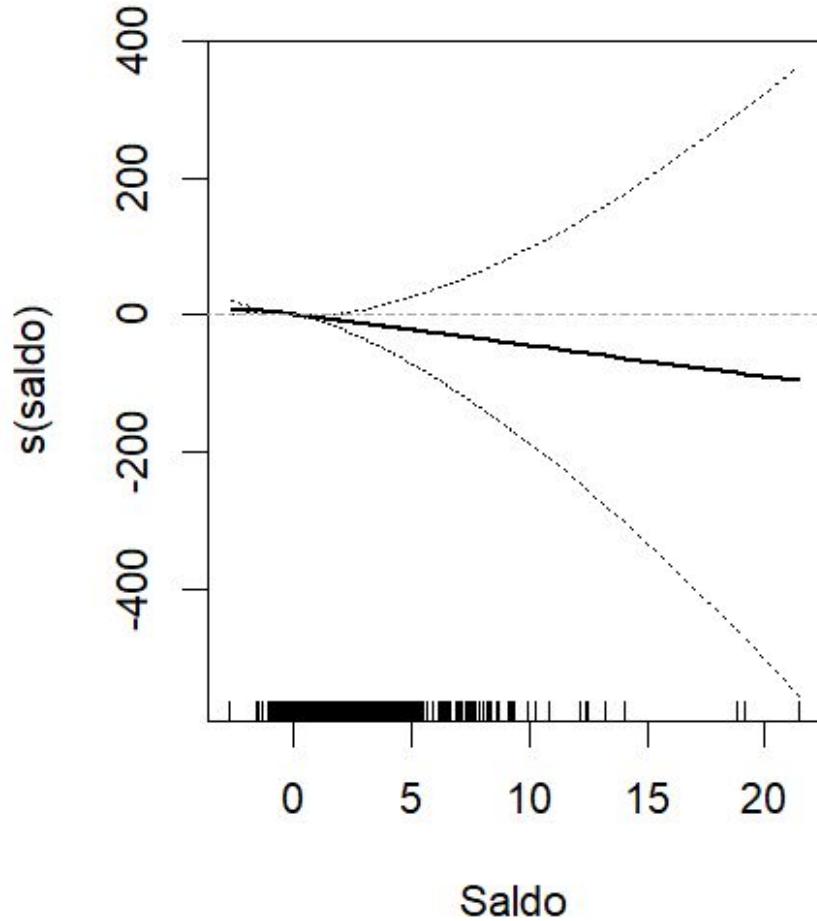
Modello	Accuracy	Recall	Precision	F1
Gradient Boosting	0.783	0.849	0.063	<b>0.117</b>
ADA Boosting	0.771	0.863	0.06	<b>0.113</b>
MARS - polspline	0.77	0.863	0.06	<b>0.113</b>
GAM	0.772	0.829	0.059	<b>0.11</b>
MARS	0.749	0.904	0.058	<b>0.109</b>
Random forest	0.787	0.753	0.057	<b>0.107</b>
GLM	0.759	0.829	0.056	<b>0.104</b>
Bagging	0.77	0.781	0.055	<b>0.103</b>
Tree	0.738	0.884	0.054	<b>0.102</b>
Lineare	0.749	0.801	0.052	<b>0.097</b>



# Interpretazione dei risultati

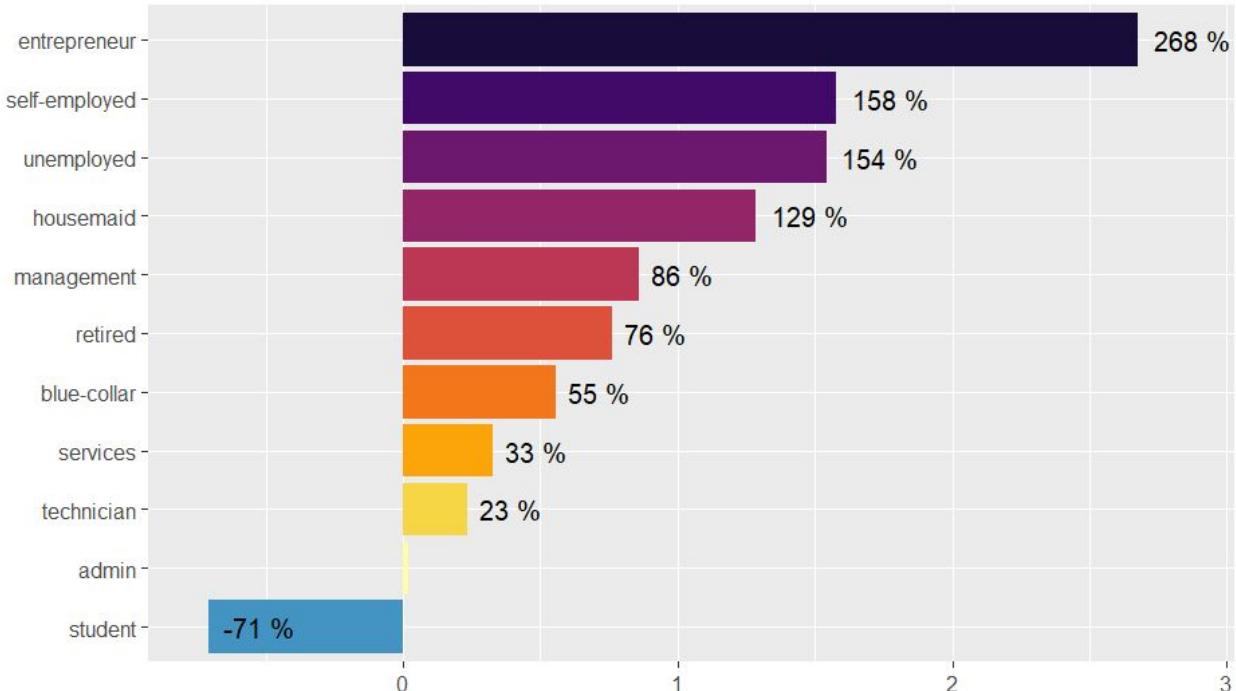


# Effetti delle variabili quantitative

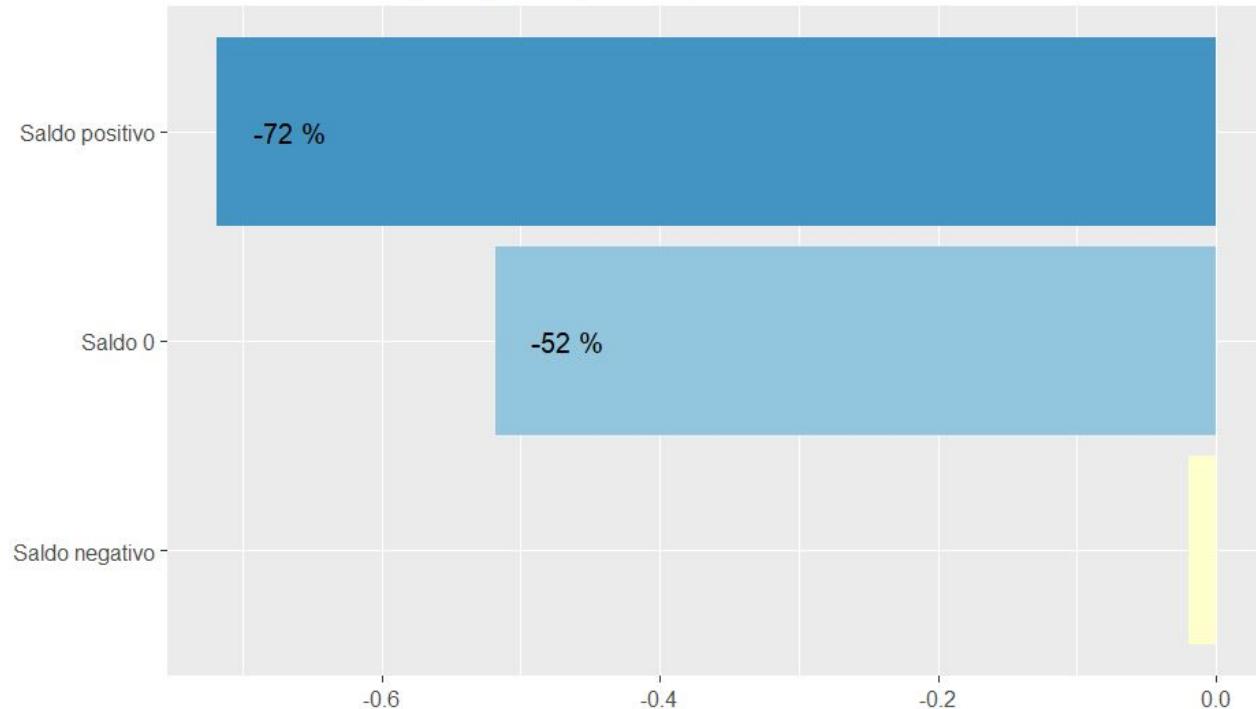


# Effetti del lavoro e del saldo (categoriale)

Effetti del lavoro sulla quota

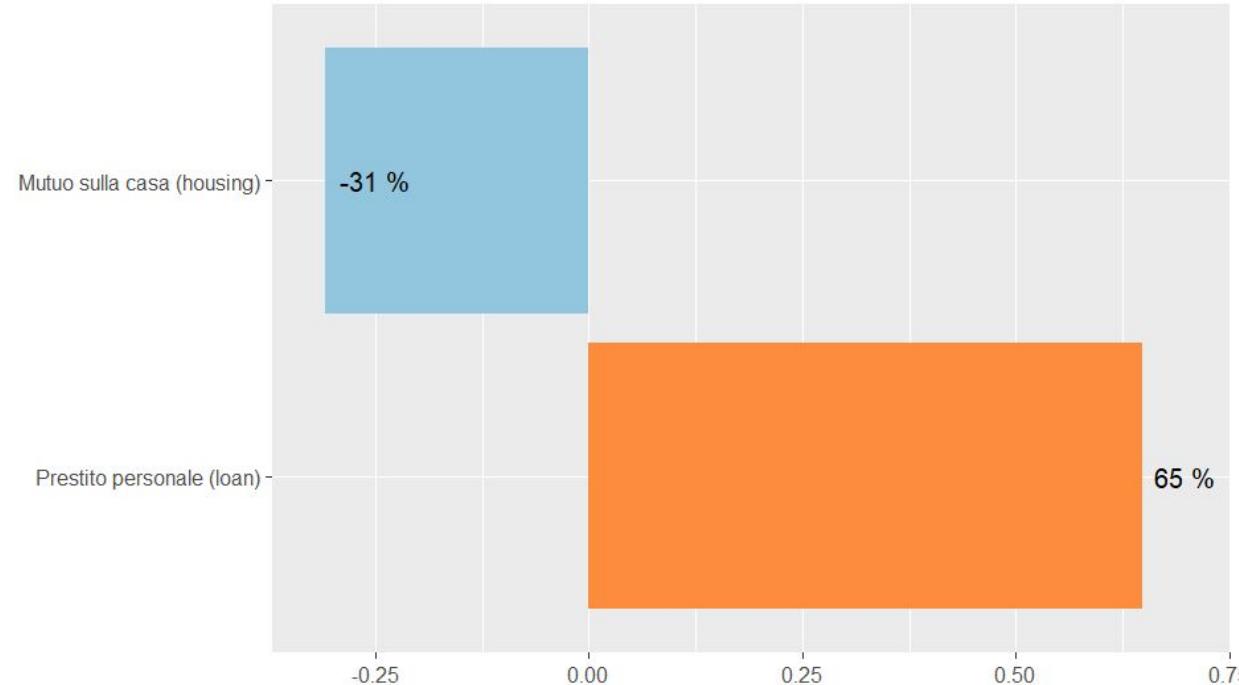


Effetti del saldo (categoriale) sulla quota

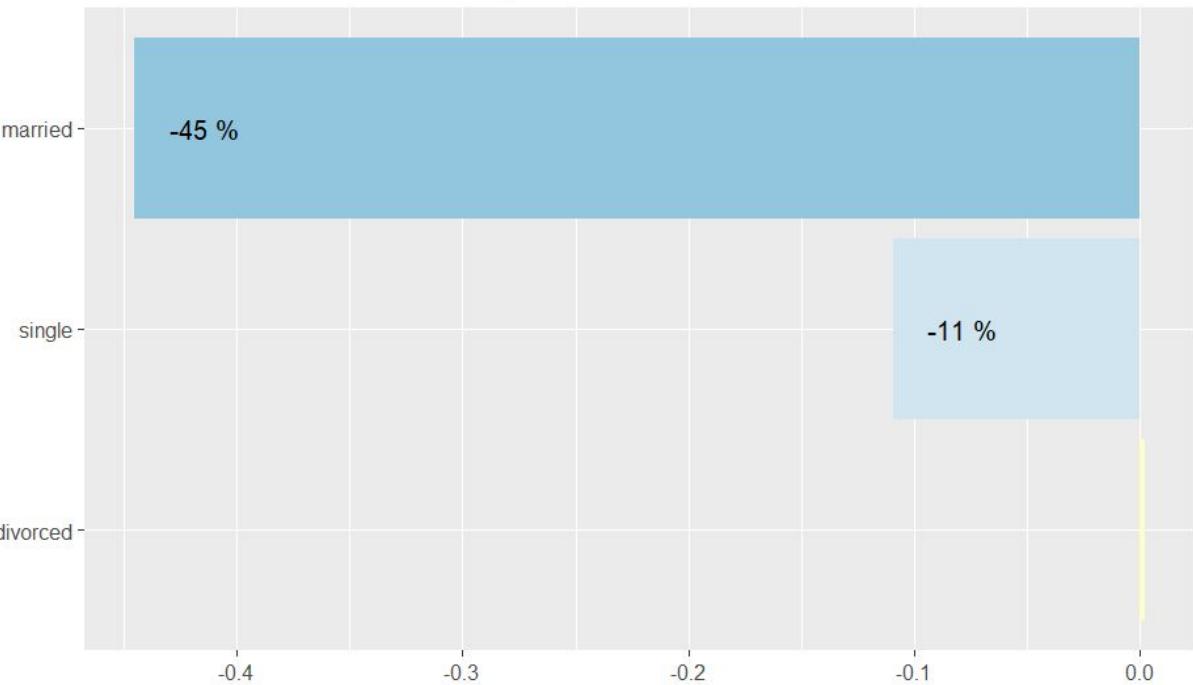


# Effetti dei prestiti e dello stato civile

Effetti dei prestiti sulla quota

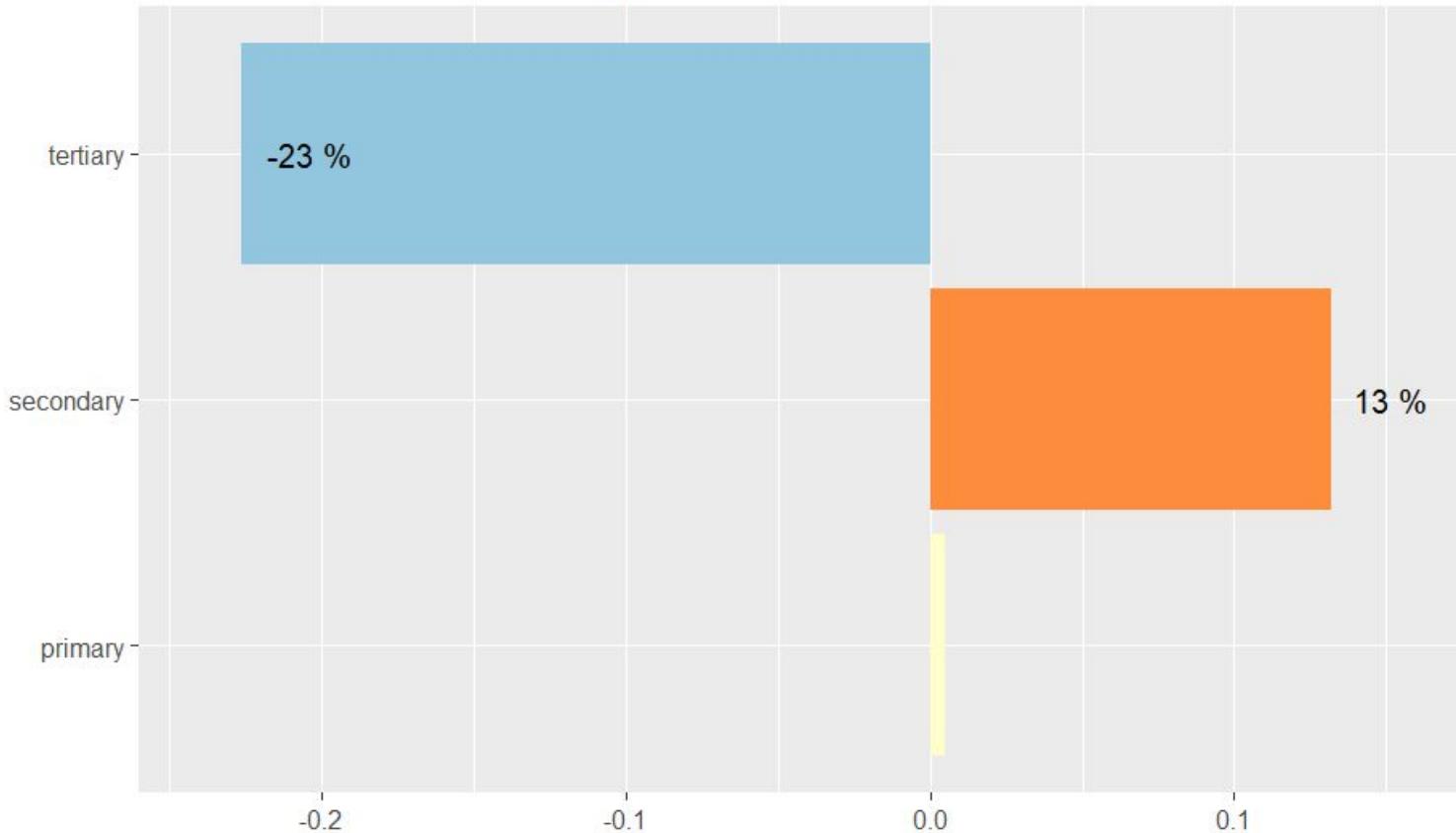


Effetto dello stato civile sulla quota



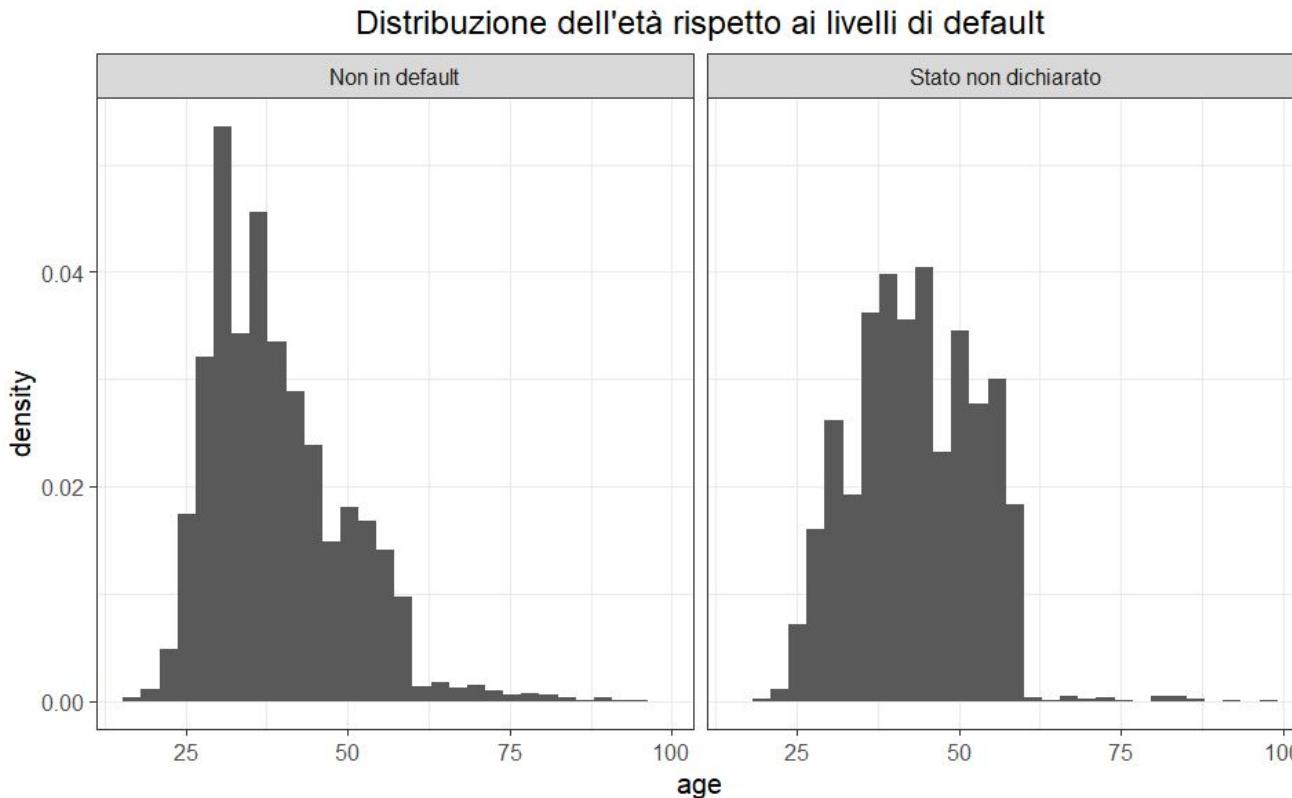
# Effetto dell'istruzione

Effetto dell'istruzione sulla quota

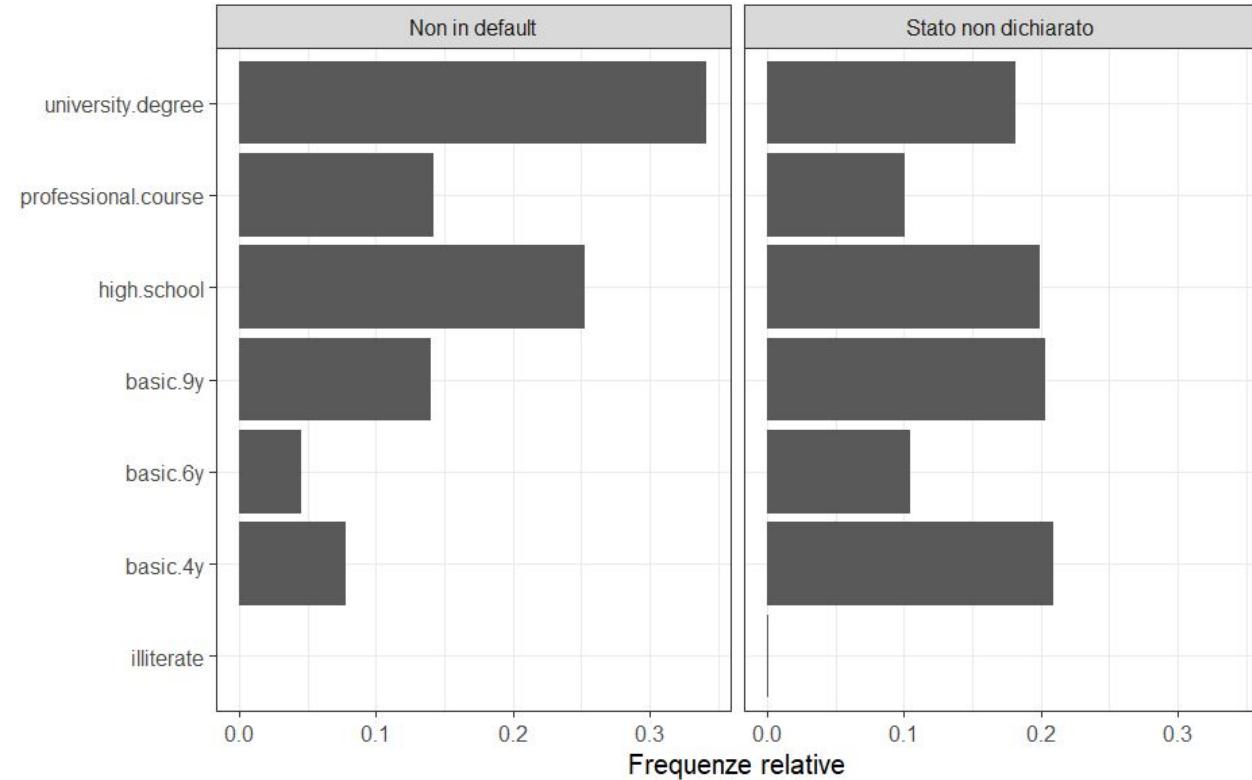
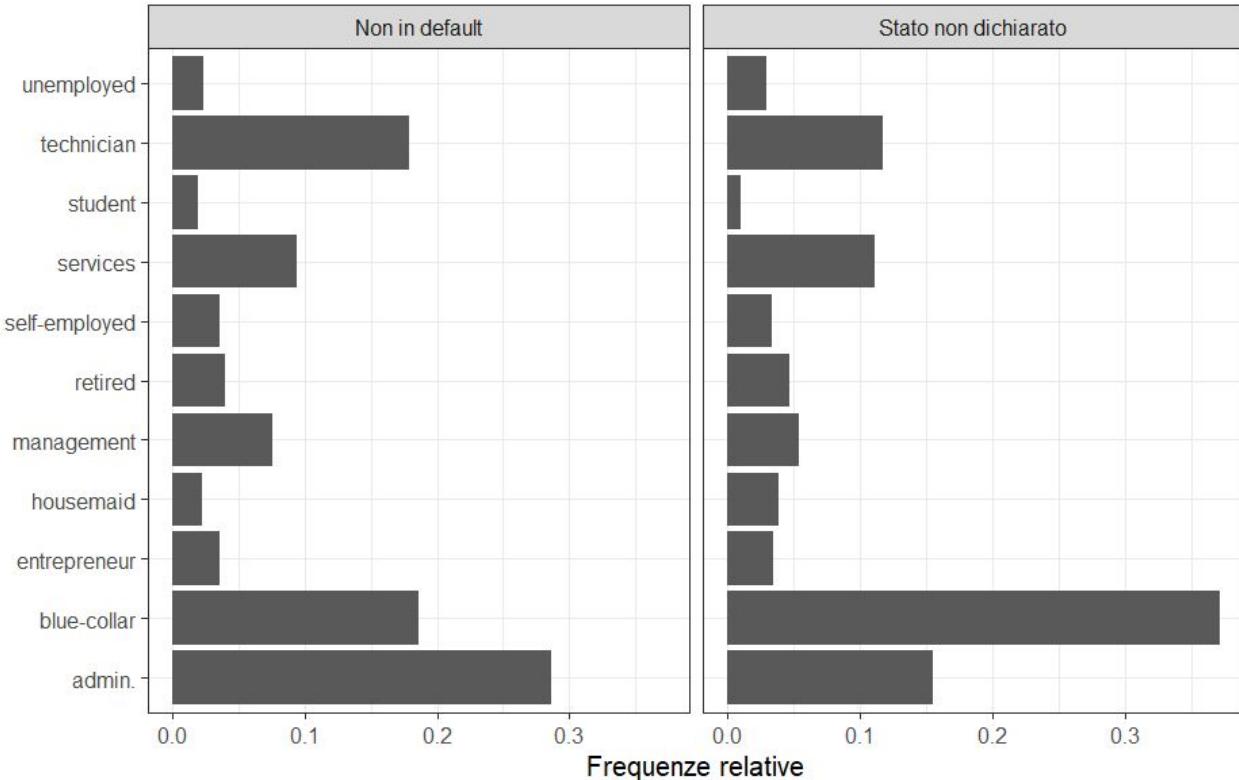


# Utilizzo dei risultati

Le conclusioni tratte da questa analisi possono essere utilizzate per chiarire la **composizione del gruppo** di individui che non dichiara lo status di default.



# Lavoro ed istruzione





# Conclusioni - Prima domanda

L'analisi ha permesso di individuare **quali consumatori sono più inclini ad accettare il prodotto offerto**, ma soprattutto è emerso come **la situazione socio-economica globale sia molto importante**.

Nei **periodi più difficili** i consumatori **tendono ad accettare più facilmente** di effettuare il deposito a m/l termine: in periodi di insicurezza economica, i clienti potrebbero cercare dei modi per assicurare il valore dei loro risparmi con un investimento.

**Consiglio per l'azienda:** investire maggiormente nelle campagne di marketing nei periodi di insicurezza economica del Paese e nei periodi in cui i tassi di interesse sugli investimenti sono più alti.

Si consiglia inoltre utilizzare il modello di gradient boosting per selezionare i clienti da contattare in future campagne di telemarketing.



# Conclusioni - Seconda domanda

Con la seconda analisi abbiamo cercato di **individuare gruppi di clienti**, sulla base delle loro caratteristiche personali, in modo da **sviluppare offerte specifiche** per ciascun gruppo.

**Consiglio per l'azienda:** provare a sviluppare un'offerta da proporre ai clienti che appartengono al terzo cluster, ovvero indirizzata ai giovani con alta istruzione e una situazione finanziaria stabile.

Il secondo gruppo, composto da clienti poco istruiti e in probabile stato di default, è da evitare, dato che probabilmente non possiedono disponibilità economiche sufficienti per investire in un deposito bancario.

Va comunque ricordato che nessun altro metodo di clustering ha prodotto dei gruppi così diversi in termini di caratteristiche e di comportamenti e quindi di affidarsi a queste conclusioni moderatamente.

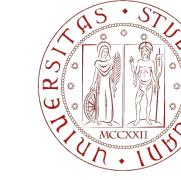


# Conclusioni - Terza domanda

Con l'ultima analisi abbiamo esplorato il fenomeno del default finanziario, cercando di capire chi sono gli **individui più a rischio**. Questo ha aiutato a chiarire meglio lo status di default degli individui che non lo hanno dichiarato.

## Consigli per l'azienda:

Si consiglia all'azienda di migliorare la metodologia di reperimento dell'informazione relativa al default finanziario, al fine di migliorare i risultati della campagna di telemarketing.



UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



Grazie per  
l'attenzione