CRAMÉR

830643 Magnacavallo Roberto

837395 Morzenti Giacomo

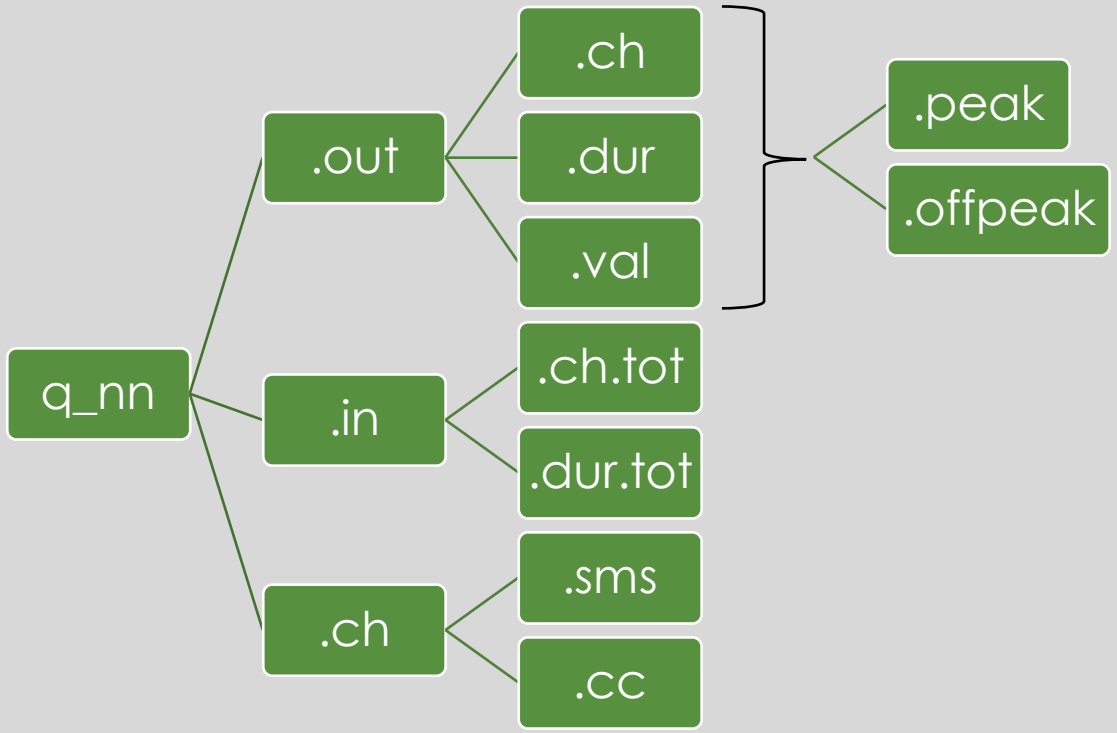
844015 Valsecchi Matteo

1. **DEFINIZIONE DEL PROBLEMA**
2. **ANALISI ESPLORATIVA**
   A. Valori mancanti-anomali
   B. Variabile VAS
   C. Potenziali problemi
   D. Features transformation end engineering
3. **APPROCCI DI REGRESSIONE**
   - Classificazione + regressione
   - Modello lineare
   - Random forest
   - Approcci basati sulla distanza
   - Ensemble model stacking

1. DEFINIZIONE DEL PROBLEMA

La direzione marketing di un'azienda di telecomunicazioni è interessata ad analizzare il comportamento di ciascun cliente relativo al suo traffico telefonico.

Si affrontare il problema di regressione nel quale si deve prevedere la durata totale delle chiamate in uscita del decimo mese di ogni cliente, in base ai dati relativi al traffico telefonico dei precedenti 9 mesi e ad alcune caratteristiche del cliente e del suo contratto telefonico.



- status
- tariff.plan
- payment.method
- gender
- age
- Activ.zone
- Activ.chan
- Vas1
- vas2

Training set	Test set
99 variabili + y	99 variabili
15310 oss.	15309 oss.

$$Err_{test} = \sum_{i=1}^{15309} [log(\hat{y}_i + 1) - log(y_i + 1)]^2$$

$$y = q10.out.dur.peak + q10.out.dur.offpeak$$

2. ANALISI ESPLORATIVA

A) VALORI MANCANTI-ANOMALI

1) Un'unica osservazione presenta il valore activ.area=0. Si considera un errore e si sostituisce con il valore moda tra i clienti «simili».

table della var activ.area

0	1	2	3	4
1	10913	9656	7217	2832

Esempio estratto

q08.in.ch.tot	q08.in.dur.tot
<int>	<int>
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0

2) Ci sono 70 osservazioni sparse nei vari mesi, (sia tra le chiamate in uscita che in entrata) con *.ch*=1 con *.dur* e *.val* (se presente) =0. Cioè osservazioni per il quale si riceve/effettua una chiamata ma con durata e valore nullo. Una prima ipotesi è che possa essere una chiamata senza risposta, ma se fosse vero è verosimile presupporre che si dovrebbe osservare (anche raramente) più chiamate perse (ad esempio .ch=2 e .dur=0), ma non si verifica mai. Perciò si decide di sostituire questi valori sospetti .ch=1 con .ch=0

3) Ci sono 12 osservazioni per il quale una variabile *.peak/.offpeak, presenta* un *.val* molto basso >0, e *.ch* e *.dur* =0. Si decide di sostituire tutti questi valori .val con 0.

Esempio estratto

q09.out.ch.offpeak	q09.out.dur.offpeak	q09.out.v…¹
<int>	<int>	<dbl>
0	0	0.0077
0	0	0.450
0	0	0.119

Esempio estratto

q04.out.ch.offpeak	q04.out.dur.offpeak	q04.out.v…¹
<int>	<int>	<dbl>
0	27	0.0036
0	46	0.0227

4) Ci sono 3 osservazioni che solo nel mese 4 e solo per variabili *.out* presentano *.ch*=0 e *.dur*>0 (con grandezza decine o centinaia) e *.val* molto basso. Poiché i valori assunti sono credibili si decide di implementare il .ch=0 con regressione lineare. (Selezionata con 15-fold cv).

5) Esistono 5 osservazioni con valori negativi, tre in *q3.in.dur.tot* e 2 in *q09.out.dur.peak. Vengono implementati con regressione lm*

Summary del mod lm usato per implementare i valori negativi di **q03.in.dur.tot**. (Selezionato minimizzando RMSE calcolato con 5 15-fold cv, forward)

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.209159   0.033836   6.182 6.49e-10 ***
q03.in.ch.tot         1.344740   0.007629 176.266  < 2e-16 ***
q.mu_no0.in.dur.tot   0.725096   0.008512  85.180  < 2e-16 ***
q.mu_no0.in.ch.tot   -1.067010   0.013176 -80.979  < 2e-16 ***
q03.out.dur.tot       0.093900   0.003238  28.995  < 2e-16 ***
q.mu_no0.out.dur.peak -0.092786   0.005129 -18.089  < 2e-16 ***
q04.in.dur.tot        0.096351   0.003919  24.584  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4932 on 17601 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9211,    Adjusted R-squared:  0.9211
F-statistic: 3.427e+04 on 6 and 17601 DF,  p-value: < 2.2e-16
```

*spoiler le variabili sono state trasformate e sono state aggiunte altre variabili.

Summary del mod lm usato per implementare i valori negativi di **q09.out.dur.peak**. (Selezionato minimizzando RMSE calcolato con 5 15-fold cv, forward)

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.084409   0.013077 312.347  < 2e-16 ***
q09.out.val.peak      1.080454   0.005220 206.988  < 2e-16 ***
tariff.plan4          0.078166   0.019305   4.049 5.16e-05 ***
tariff.plan6          0.140676   0.010697  13.150  < 2e-16 ***
tariff.plan7          0.112710   0.008830  12.765  < 2e-16 ***
tariff.plan8         -0.105189   0.008530 -12.331  < 2e-16 ***
vas1                  0.131249   0.004269  30.747  < 2e-16 ***
vas2                  0.049976   0.007987   6.257 3.98e-10 ***
q08.out.ch.offpeak -0.032683   0.002371 -13.783  < 2e-16 ***
q08.out.dur.peak      0.061422   0.002094  29.337  < 2e-16 ***
q08.out.val.peak     -0.123740   0.004191 -29.523  < 2e-16 ***
q09.out.ch.peak       0.153625   0.005730  26.809  < 2e-16 ***
q07.in.ch.tot        -0.023394   0.002111 -11.082  < 2e-16 ***
q07.out.dur.peak      0.076458   0.002610  29.291  < 2e-16 ***
q07.out.ch.peak      -0.134246   0.005199 -25.822  < 2e-16 ***
q01.ch.sms            0.021712   0.002771   7.837 4.80e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2875 on 25983 degrees of freedom
Multiple R-squared:  0.9682,    Adjusted R-squared:  0.9682
F-statistic: 5.269e+04 on 15 and 25983 DF,  p-value: < 2.2e-16
```

B) VARIABILE VAS

Le variabili vas1 e vas2 hanno entrambe 2 livelli: «Y,N» ma non si osservano mai =Y congiuntamente. Quindi se una è =Y l'altra sarà necessariamente =N. Perciò è possibile passare da una codifica in 2 variabili a due livelli ad una codifica di un'unica variabile VAS a 3 livelli senza perdita di informazione (0 se vas1 e vas2 sono =N, 1 se vas1=Y e vas2=F, 2 se vas1=F e vas2=T).

table congiunto di vas1 e vas2

	N	Y
N	20938	1828
Y	7853	0

$$vas = \begin{cases} 0 & \text{if} \quad vas1 = vas2 = N \\ 1 & \text{if} \quad vas1 = Y \\ 2 & \text{if} \quad vas2 = Y \end{cases}$$

Da una più attenta analisi si è però evidenziato che le variazioni contrattuali non si sono mai osservate congiuntamente non perché sono condizioni mutualmente esclusive, ma perché nel caso in cui un cliente attivasse entrambe le variazioni contrattuali verrebbe registrato con due osservazioni identiche ma con vas diverso.

Quindi non tutte le osservazioni si riferiscono a clienti diversi!

Esistono osservazioni uguali con vas diverso:

- Coppie di oss con: vas=1 e vas=2
- Coppie di oss con: vas=0 e vas=1 oppure vas=2
- Terna di oss con: vas=0, vas=1, vas=2

Riuscire ad identificare le osservazioni appartenenti allo stesso cliente permette di:

- prevedere con certezza la y di xxx osservazioni nel test set (poiché sono «copie» di osservazioni del training set)
- Nel caso in cui più osservazioni del test set si identificano come appartenenti allo stesso cliente, e nel caso in cui a seguito della regressione verranno attribuiti valori diversi, sarà possibile combinare questi valori (ad esempio con una media) e assegnare alle osservazioni la stessa stima di y

Esempio estratto

tariff.plan	age	vas	q08.in.ch.tot	q09.out.val.peak	q09.out.val.offpeak	y
7	30.98	2	174	205.5767	0	*NA*
7	30.98	1	174	205.5767	0	2139

Osservazione del test n. 35

Osservazione del training n. 14602

```
> all(select(test[56,],-vas,-y)==select(train[14602,],-vas,-y))
[1] TRUE
```

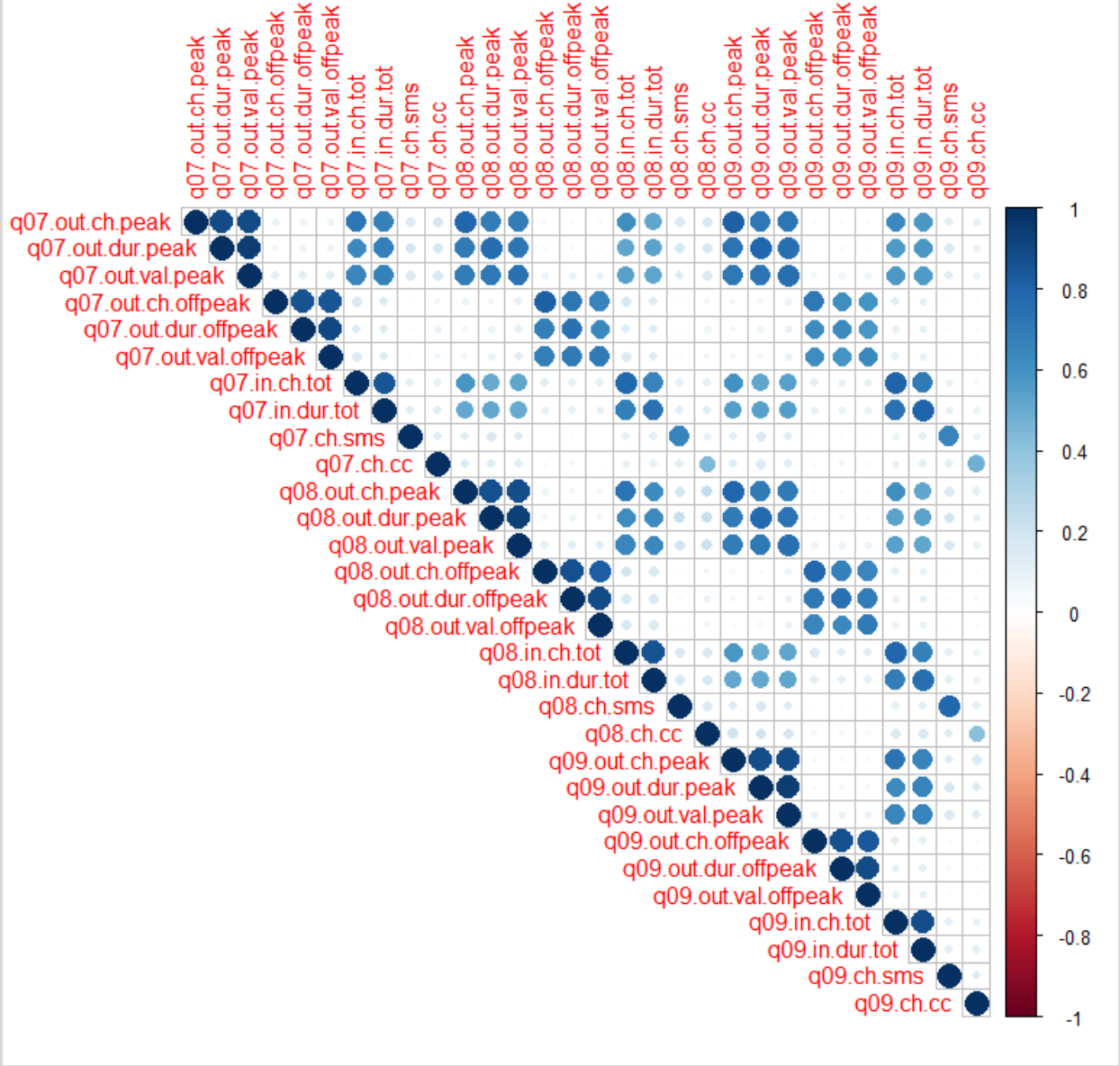
C) POTENZIALI PROBLEMI

- PROBLEMI DI DISTRIBUZIONE: La variabile Y e le variabili q hanno forti problemi di distribuzione caratterizzati da una elevata varianza e la presenza di una grande quantità di outlier. La Y presenterà anche un problema di bimodalità in 0.

- PROBLEMI DI CORRELAZIONE E INF. RIDONDANTI: Esiste un'elevata correlazione tra variabili anche per la loro stessa natura, sia per le variabili che si riferiscono alla stessa tipologia di consumo del mese (il numero di chiamate di un mese sarà altamente correlato alla loro durata totale e al loro valore), sia temporale (il numero di chiamate di una tipologia in un mese sarà correlato con il numero di chiamate della stessa tipologia nei vari mesi, con una correlazione sempre più «debole» più ci si allontana temporalmente).

- OSSERVAZIONI CON TUTTE LE VAR q=0: Esiste un gruppo folto di clienti che per tutti i 9 mesi non chiamano e non ricevono chiamate e messaggi. La maggior parte di questi clienti hanno Y=0, ma alcuni di loro Y!=0. La quasi totalità di questi clienti sarà molto difficili da identificare.

Matrice di correlazione delle variabili relative agli ultimi 3 mesi



D) FEATURES TRANSFORMATION END ENGINEERING

- Per risolvere il problema di distribuzione di y e delle variabili q, si decide di trasformare tutte queste variabili con la trasformazione logaritmica. Prima di applicare il logaritmo si somma 1 per la presenza degli 0. Questa trasformazione risolve questi problemi ed evidenzia la forte bimodalità di y in 0.

$$X = log(X + 1)$$

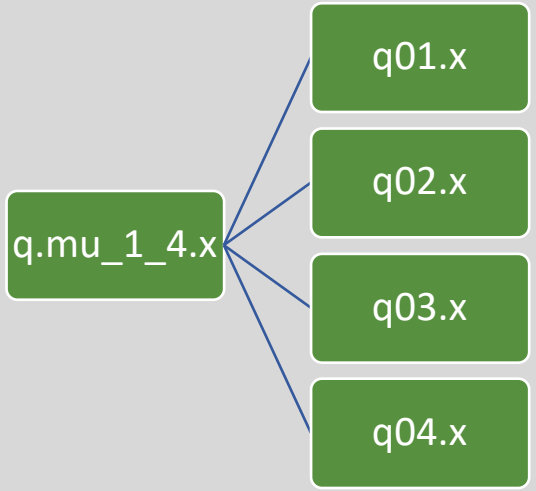
$$Err_{test} = \sum_{i=1}^{15309} [log(\hat{y}_i + 1) - log(y_i + 1)]^2 =$$

$$= \sum_{i=1}^{15309} (\hat{x}_i - x_i)^2 = SSE_{\hat{x}}$$

OSS: la funzione di errore del test, equivale all' SSE della previsione della y trasformata sul test. Quindi minimizzare MSE di previsione della y trasformato sul test, sarà equivalente a minimizzare la funzione di errore.

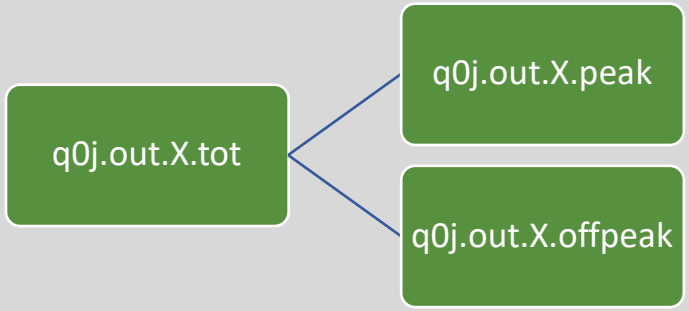
$$: \; x_i = log(y_i + 1) \quad \forall i = 1 \dots 15309$$

- Per gestire in parte il problema di correlazione e di ridondanza delle informazioni si sono create diverse nuove variabili con il fine di provare ad estrarre e concentrare dell'informazione in «poche» variabili e se possibile ridurre la dimensionalità.



Si introduce una variabile per ogni «tipologia», calcolata come media delle prime 4 variabili di tale «tipologia». Con lo scopo di mantenere parzialmente le informazioni dei primi 4 mesi e concentrarla in un'unica variabile per «tipo». L'introduzione di queste variabili ha permesso la rimozione di tutte le variabili dei primi 4 mesi.
Ad esempio, per la «tipologia» .out.ch.peack:

$$q.mu\_1\_4.out.ch.peak = \frac{1}{4}\sum_{j=1}^{4} q.0j.out.ch.peak$$



Si sono aggiunte le variabili .tot riguardo il consumo in uscita, aggregando le variabili .peak e .offpeak della stessa «tipologia» (essendo y la durata totale di tutte le chiamate in uscita). Ad esempio:

$$q0j.out.ch.tot = q0j.out.ch.peak + q0j.out.ch.offpeak$$

q.mu_no0.X

Per ogni «tipologia» si è aggiunta una variabile calcolata come la media dei valori !=0 delle variabili di tipologia X. Ad esempio la rispettiva variabile per la «tipologia» .out.ch.peak :

$$q.mu\_no0.out.ch.peak = \frac{1}{\sum_{j=1}^{9} \mathbb{I}\{q.0j.out.ch.peak > 0\}} \sum_{j=1}^{9} q.0j.out.ch.peak$$

n0.X

Si aggiunge una variabile per ogni «tipologia» X che conta quante variabili di «tipo» X sono =0
Ad esempio, per la «tipologia» .out.ch.peack :

$$n0.out.ch.peak = \sum_{j=1}^{9} \mathbb{I}\{q0j.out.ch.peak = 0\}$$

n0.mesi

Si aggiunge un'unica variabile che conteggia quanti mesi il cliente non ha ne effettuato chiamate in uscita, ne ha ricevuto chiamate e messaggi, cioè il numero di mesi nel quale tutte le variabili q sono =0. Cioè:

$$n0.mesi = \sum_{j=1}^{9} \mathbb{I}\{\sum_{x \in X} \mathbb{I}\{q0j.x > 0\} = 0\}$$

Si sono rimosse tutte le variabili relative alle chiamate verso i servizio clienti (.cc) e tutte quelle relative ai messaggi ricevuti (.sms)

3. APPROCCI DI REGRESSIONE

CLASSIFICAZIONE + REGRESSIONE

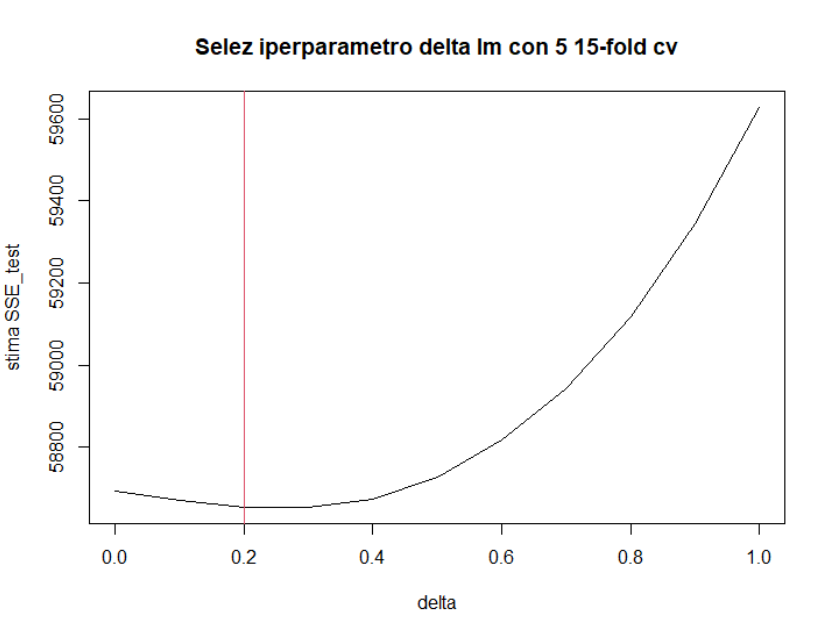
- Per affrontare il problema di bimodalità della y la prima idea è stata quella di effettuare prima una classificazione in modo da indentificare le osservazioni =0 e quelle >0, per poi effettuare una regressione solo per le osservazioni identificate come >0.
- Ma questo approccio è stato abbandonato poiché i primi risultati hanno mostrato un errore più alto confronto i metodi di sola regressione. Ciò viene principalmente giustificato da:
    I. La difficoltà nel discriminare le osservazioni (ad esempio la presenza delle osservazioni con tutte le variabili q=0)
    II. La struttura dell'errore (quadratico)

Per questi motivi, nonostante una accuracy «buona» i falsi 0 portano ad un errore maggiore confronto il guadagno nell'identificare le osservazioni con y=0. Invece, sembra essere preferibile un metodo che attribuisca valori di y anche molto «piccoli» invece che nulli in tutti i casi.

Modello lineare

Si è selezionato il modello migliore tra i due modelli ottimi selezionati con il metodo backward e il metodo forward minimizzando RMSE stimato con 5 15-fold cv. Cioè il modello selezionato con il metodo backward con 61 variabili e SSE stimato di circa 59k.

SSE_test stimato con 5 15-fold cv	58653
Delta	0.2
n. Variabili	61



Poiché il modello restituisce anche previsioni negative, si è definito e selezionato (con cv) il parametro delta:

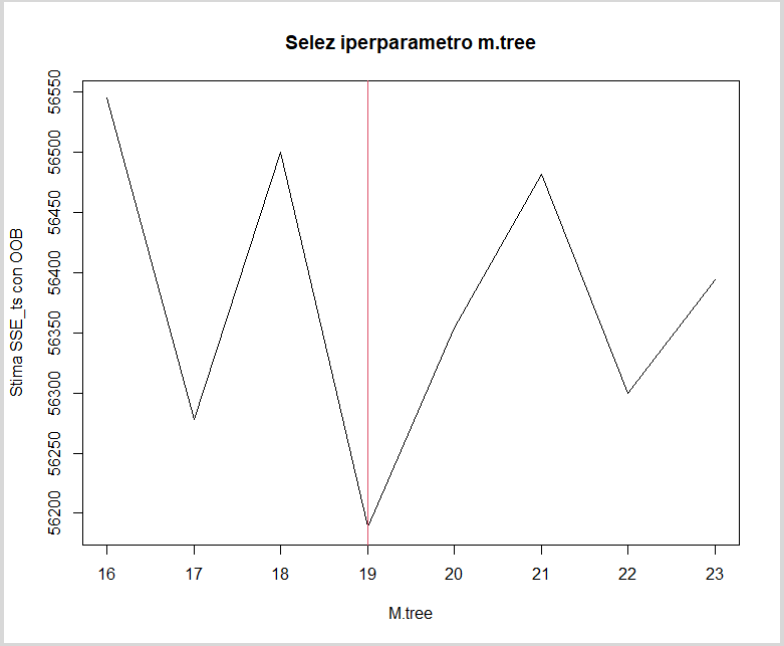
$$\hat{y} = \begin{cases} \delta & \text{if } \hat{y} < \delta \\ \hat{y} & \text{otherwise} \end{cases}$$

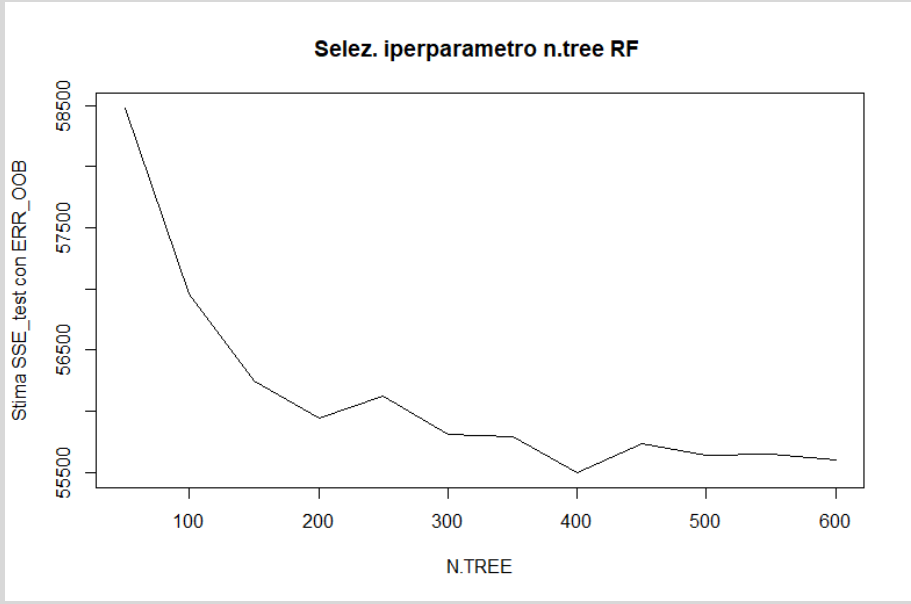
Random forest

Le variabili categoriali sono state ricodificate, utilizzando la y media della classe.

Si è selezionata la rf ottima con m.try=19m n.tree=350 con SSE_test stimato con OOB circa 55k

SSE_test stimato con OOB	55763
m.try	19
n.tree	300

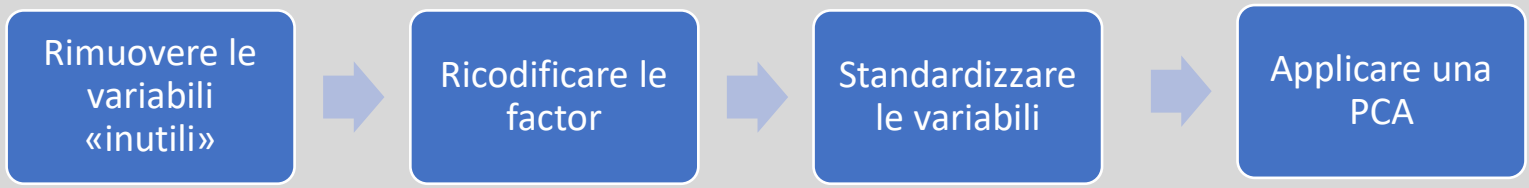




METODI BASATI SULLA DISTANZA

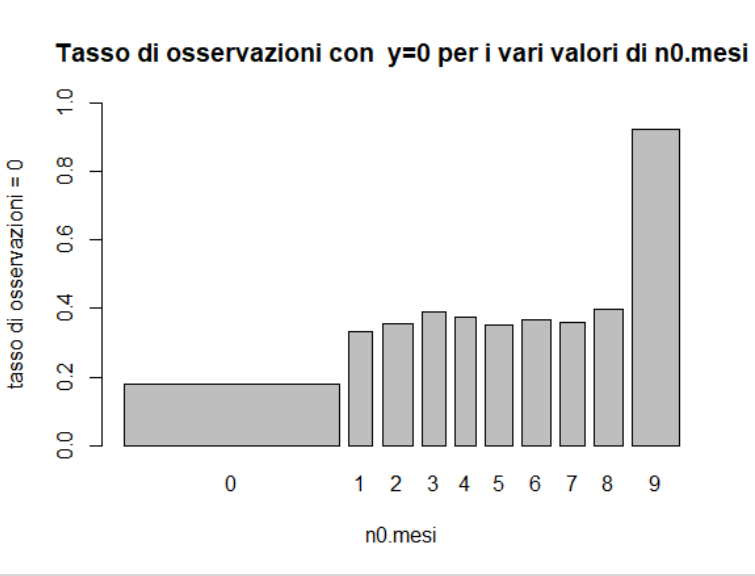
Per poter applicare dei metodi di regressione basati sulle misure di distanza bisogna affrontare diversi problemi:

- La presenza di variabili «inutili»
- La presenza di variabili correlate
- La forte diversità tra varianze
- La presenza di variabili categoriali



1. Si rimuovono tutte le variabili che sono combinazioni lineari di altre variabili (cioè le variabili aggiunte q.mu_no0, q0j.out.x.tot)
2. Si rimuovono le variabili che conteggiano i mesi nulli (n0.X) ad eccezione di n0.mesi che viene ricodificata come factor in in 3 livelli

$$x = \begin{cases} x & \text{if } x \in \{0, 9\} \\ 1 & \text{otherwise} \end{cases}$$



3. Si codificano tutte le variabili categoriali usando la y media della rispettiva classe

$$x = \frac{1}{\sum_{i=1}^{1310} \mathbb{I}\{X_i = x\}} \sum_{i=1}^{15310} y_i \cdot \mathbb{I}\{X_i = x\}$$

In questo modo si ottengono 57 variabili

Problema della Standardizzazione

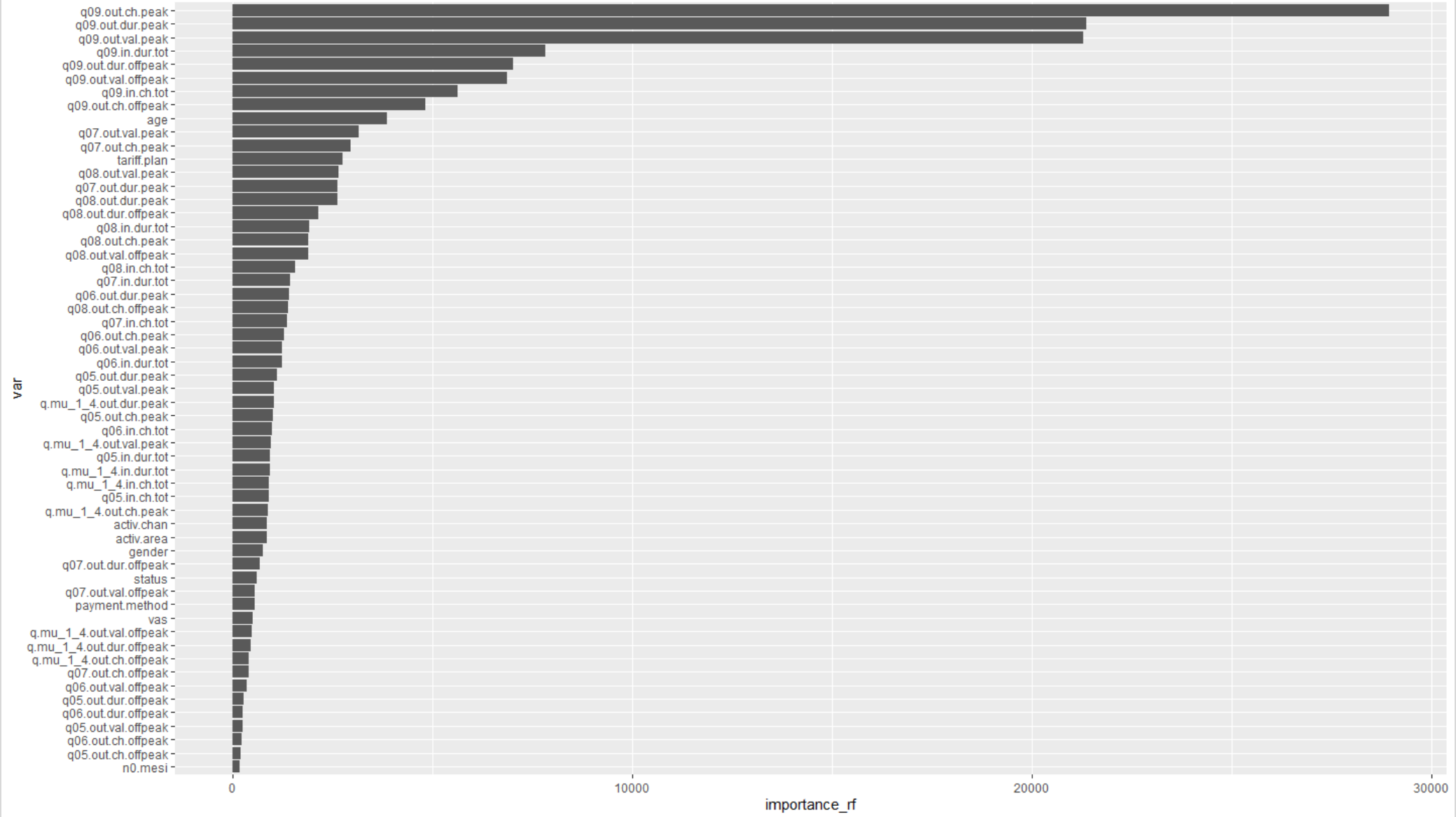
Standardizzare le variabili prima della PCA è fondamentale per la grande differenza di varianza tra le variabili. Ma applicando una standardizzazione si imporrebbe lo stesso peso a tutte le variabili per la formazione delle componenti. Perciò si decide di fissare la varianza delle variabili secondo un criterio di importanza, in modo che le variabili più importanti avranno una varianza maggiore e quindi un peso maggiore nella formazione delle componenti.

$$X = \frac{X - mu_X}{\sigma_X} \cdot \sqrt{w_x} \qquad : w_x \equiv \text{varianza desiderata di X}$$

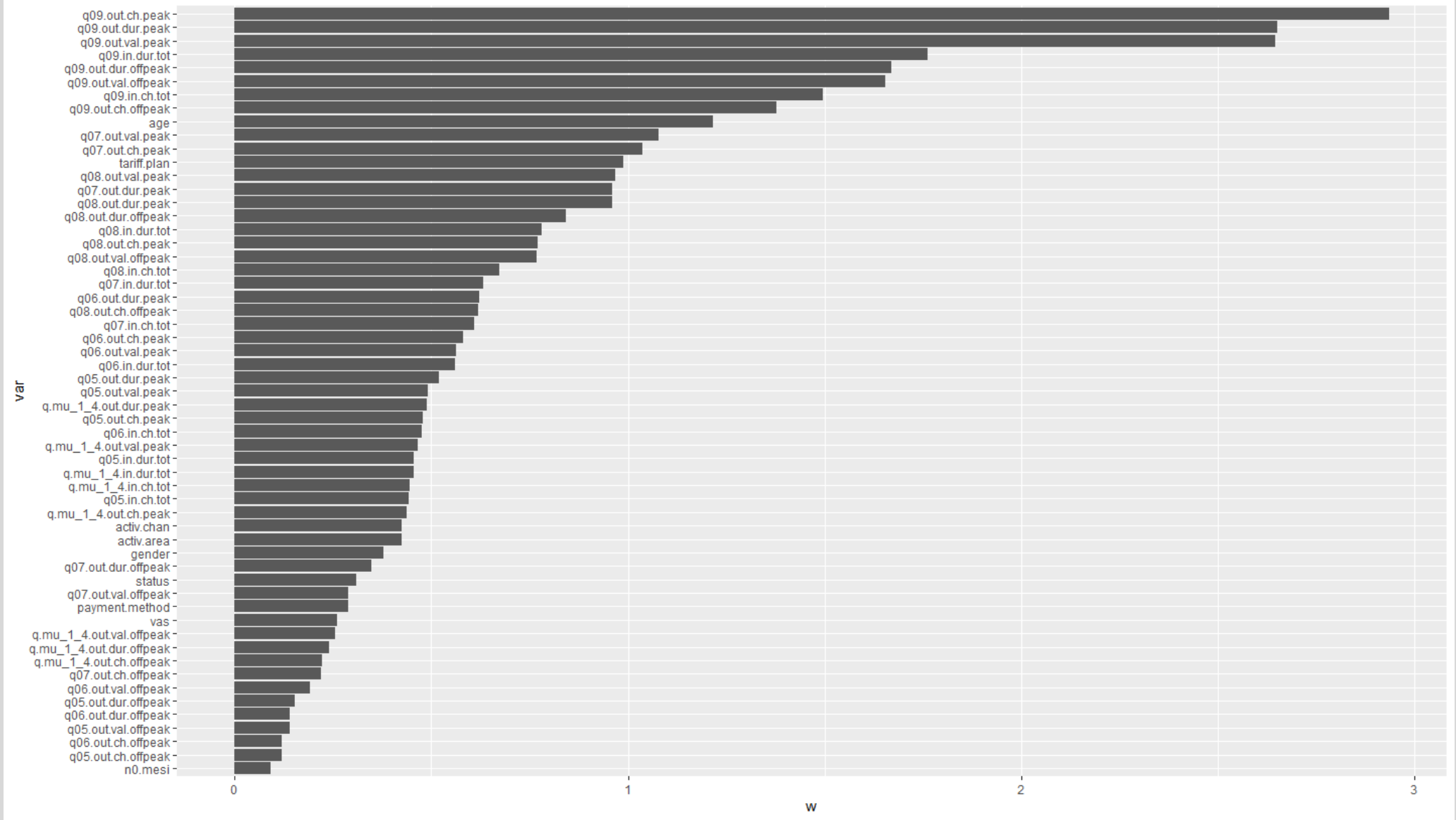
I pesi si sono ottenuti come trasformazione dei valori di importanza calcolati con la random forest

$$w_x = log(\frac{p_x}{\sum_{j=1}^{57} p_{x_j}} + 1) \qquad : p_{xj} \equiv \text{importanza di} X_j$$

Coefficienti di importanza della random forest

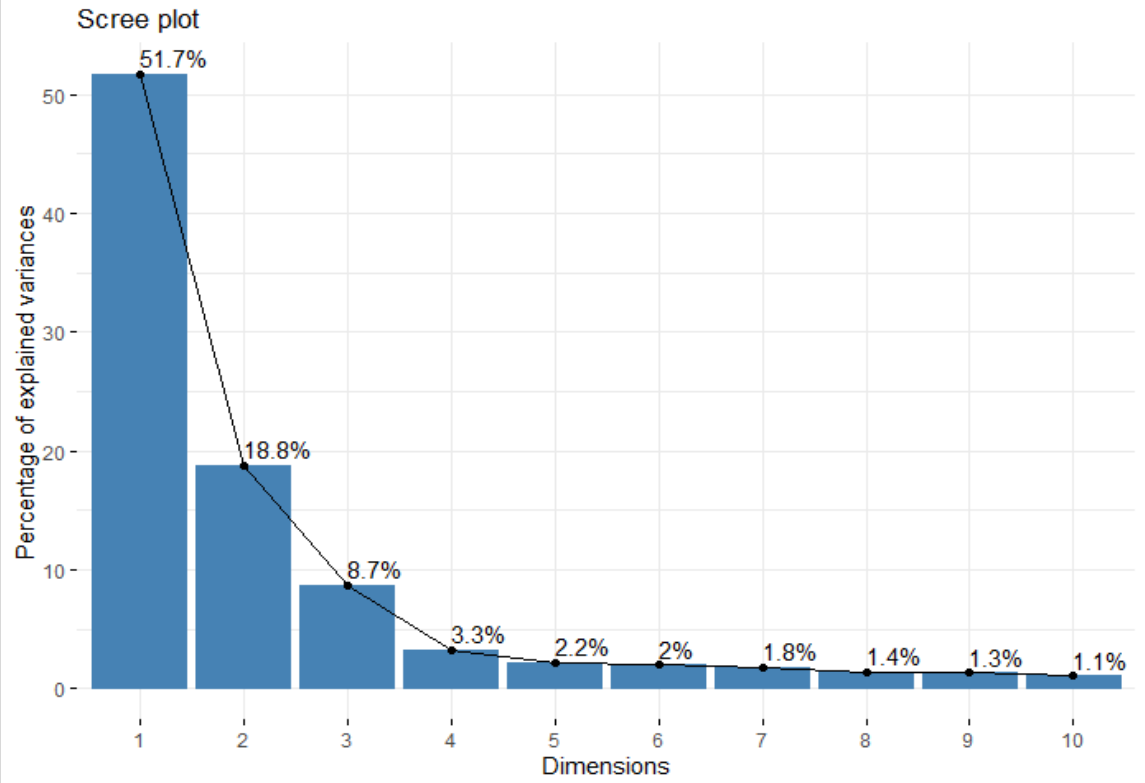


Pesi calcolati



Si è applicata una PCA e sono state selezionate le prime **26** componenti con una varianza spiegata del **99%** (le variabili erano 57)

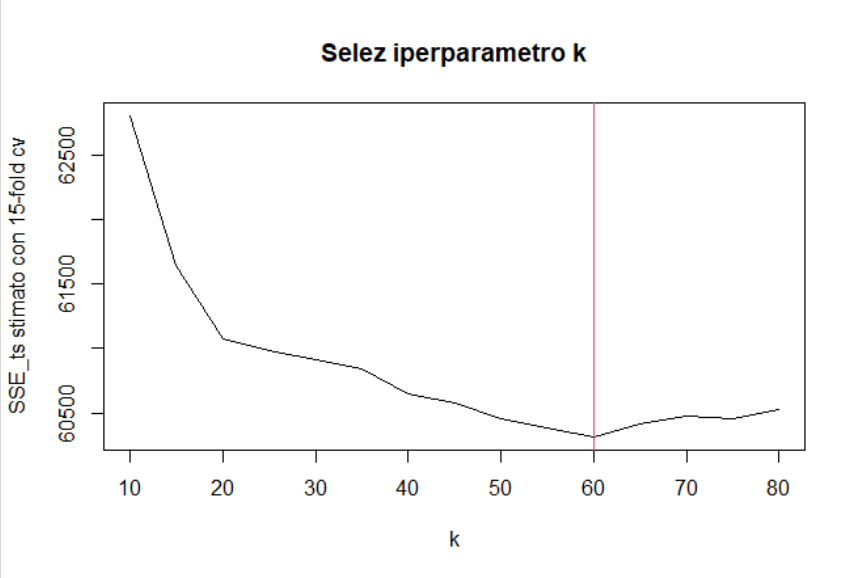
Percentuale di variabilità spiegata dalle componenti



Knn

Si è selezionato k=60 con 15-fold cv, con SSE_ts stimato di circa 60k

SSE_test stimato con 15-fold cv	60398
k	60



Si è stimata anche una SVM lineare con SSE_test stiamta di circa 63k

ENSEMBLE MODEL STACKING

I pesi sono stati stimati con 2 10-fold cv.

Successivamente si è stimato SSE_test con 2 10-fold cv di tutti i modelli e di tutti gli ensemble costruiti con i pesi stimati.

Si è deciso di scegliere l'ensemle costituito da RF e LM.

Matrice dei pesi stiamti

reg	MSE_ts	w.lm	w.rf	w.knn	w.svm
<chr>	<db1>	<db1>	<db1>	<db1>	<db1>
lm	4.02	NA	NA	NA	NA
rf	3.85	NA	NA	NA	NA
knn	4.14	NA	NA	NA	NA
svm	4.32	NA	NA	NA	NA
lm+rf	3.80	0.292	0.719	0	0
lm+rf+knn	3.80	0.330	0.740	-0.0598	0
lm+rf+svm	3.79	0.452	0.756	0	-0.186
lm+rf+knn+svm	3.79	0.451	0.755	0.00530	-0.190

SSE_test stimati

reg	MSE_ts	SSE
<chr>	<db1>	<db1>
lm	4.02	58633.
rf	3.85	56105.
knn	4.14	60383.
svm	4.32	62971.
lm+rf	3.81	55508.
lm+rf+knn	3.81	55497.
lm+rf+svm	3.80	55387.
lm+rf+knn+svm	3.80	55386.

GRAZIE PER L'ATTENZIONE