Michela Ropele

Data Mining: Clima

1. Quesito di ricerca

Affrontare l'emergenza climatica è oggi una delle sfide più urgenti e rilevanti per la nostra società.

Negli ultimi anni, numerosi Paesi hanno discusso e implementato politiche per mitigare il cambiamento climatico. Spesso tali politiche richiedono risorse e sacrifici da parte dei cittadini, suscitando frequentemente opinioni contrastanti: c'è chi non condivide gli obiettivi e chi li condivide ma non necessariamente concorda con le misure adottate per raggiungerli. C'è inoltre chi nega la presenza di un cambiamento climatico.

All'interno del progetto "Trust in Science and Science-Related Populism" (TISP), è stata condotta un'indagine in diversi paesi per misurare la percezione su queste tematiche.

Il questionario proposto agli intervistati includeva misure approfondite sugli atteggiamenti nei confronti del cambiamento climatico e sul sostegno alle politiche ambientali.

Contestualmente sono state anche raccolte opinioni relativamente alla fiducia negli scienziati, agli atteggiamenti verso la scienza, alle percezioni del ruolo della scienza nella società, all'uso dei media scientifici e ai comportamenti comunicativi, oltre ad opinioni politiche e religiose e alle caratteristiche demografiche degli intervistati.

Il file core-questionnaire_english.pdf contiene i dettagli del questionario e delle domande rivolte ai cittadini. Il file clima.csv contiene le risposte di (circa) 70000 individui alle 140 domande.

È di interesse valutare come il livello di supporto alle politiche per la lotta all'emergenza climatica vari tra le diverse nazioni e in che modo tale percezione sia influenzata dal grado di fiducia dei cittadini nella scienza e negli scienziati, al netto delle altre variabili di controllo disponibili nell'indagine.

In particolare, le domande contrassegnate con il prefisso "CLIM_POLSUPPORT_" misurano il livello di supporto (1:nessuno, 2: moderato, 3: elevato) espresso da ciascun partecipante nei confronti di alcune politiche per contrastare il cambiamento climatico (tasse sui combustibili, investimenti nei trasporti pubblici, promozione di energie sostenibili, salvaguardia delle aree forestali e tassazione degli alimenti ad alto impatto ambientale).

Questi item possono essere aggregati in uno (o più) indicatore che esprima il livello complessivo di adesione a tali politiche.

2. Dati

Il dataset presenta inizialmente 68054 osservazioni e 141 variabili. Le unità statistiche sono i singoli questionari. Non ci sono osservazioni ripetute.

La risposta è composta da 5 variabili ordinali con valori 1, 2 e 3, più una categoria neutra (4). Per ottenere una risposta unidimensionale, le variabili vengono aggregate tramite la media, escludendo le risposte con valore neutro (4) o mancanti. Per esempio, se un'osservazione presenta 3 risposte con modalità ordinale (1, 2 o 3), una neutra (pari a 4) e un valore mancante, la loro media complessiva sarà solamente la media dei primi tre valori. La nuova variabile viene normalizzata nell'intervallo (0,1). Le variabili originarie vengono rimosse per evitare collinearità.

L'obbiettivo dell'analisi è capire come la risposta varia tra le varie nazioni e cosa la influenza.

Eseguo le seguenti operazioni di pulizia:

- Rimuovo l'identificativo di riga;
- Rimuovo tutte le variabili relative alla metrica di compilazione del questionario, ad eccezione del paese di appartenenza;
- Le variabili sull'orientamento politico (DEM_POL_conservative, DEM_POL_right) presentano valori mancanti; a questi viene assegnata la modalità "99 = non so". Pur avendo cause diverse, le mancanze vengono trattate in modo uniforme ai fini dell'analisi;
- Per tutte le altre variabili con valori mancanti, elimino le osservazioni corrispondenti, poiché la numerosità campionaria è sufficientemente elevata da non compromettere la rappresentatività;
- Rimuovo le variabili TRUST_OPEN e BENEFIT_OPEN perché presentano delle risposte testuali e non sono utili ai fini dell'analisi.
- Ci sono 3 viabili relative al reddito che restituiscono la stessa informazione. Tengo DEM_INCOME_USD;
- La variabile COUNTRY_NAME presenta 5 modalità senza osservazioni: Albania, Brazil, Finland Mexico, Perù. Rimuovo queste modalità rimanendo con 61 modalità.

Ora il dataset presenta 56693 osservazioni e 103 variabili.

3. Operazioni preliminari

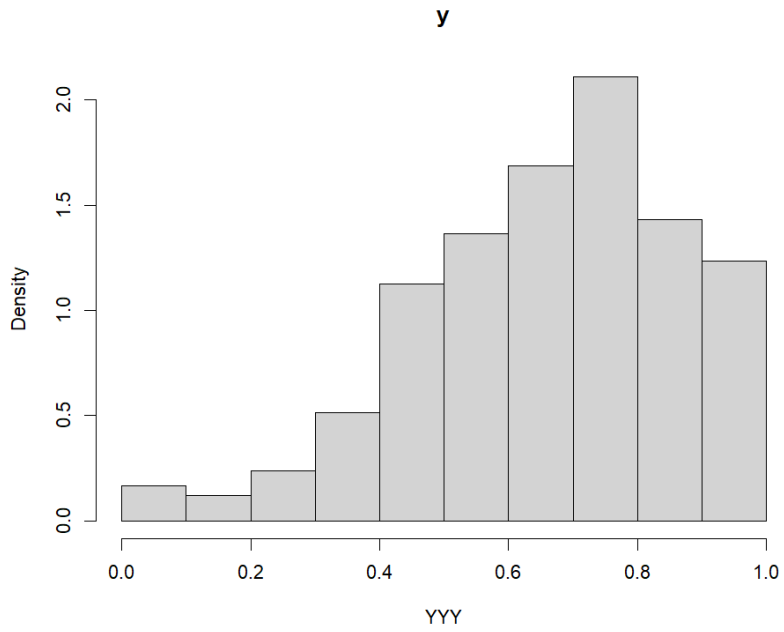
Suddivido il dataset in stima e verifica (70% e 30%), dividendo ulteriormente il dataset di stima in stima e convalida per la selezione dei parametri di regolazione (2/3 e 1/3 delle osservazioni rispettivamente).
Tale procedura è appropriata poiché le osservazioni non hanno una componente temporale.

Nell'insieme di verifica non ci sono variabili con correlazione superiore (in valore assoluto) a 0.95, ma ci sono 25 coppie di variabili con correlazione compresa tra 0,70 e 0,80. Dovrò tenere conto di questa cosa.

Viene, inoltre, eseguita la standardizzazione separata delle variabili quantitative nei due sottoinsiemi per rendere i coefficienti confrontabili tra loro in termini di una deviazione standard (ad esempio nel Modello Lineare senza interazioni), oppure evitare che alcune variabili dominino l'analisi in modelli sensibili alla scala (ad esempio nella PCA).

4. Variabile risposta

La risposta presenta la seguente distribuzione marginale. Si osserva che è asimmetrica con una coda a sinistra ed una media in corrispondenza di circa 0,75.



5. Prestazioni del modello

Per valutare le diverse previsioni, utilizzo le metriche:

- MSE che penalizza molto gli errori grandi e poco quelli piccoli;
- Perdita logaritmica, adatta per valutare modelli con output di valori compresi tra 0 e 1. La formula è la seguente:

$$Log - loss(y_i) = -1/n * sum\big(y_i * log(p_i) + (1 - y_i) * log(1 - p_i)\big)$$

  Dove le y_i sono i valori osservati della risposta e p_i le previsioni ottenute dal modello. Più piccolo è il valore osservato, minore sarà la sua trasformata logaritmica. Dunque, valori osservati e predetti coerenti producono una perdita bassa, mentre previsioni errate o troppo estreme comportano invece una penalizzazione maggiore.

Per la selezione dei modelli utilizzerò l'AIC (ovvero la verosimiglianza penalizzata dalla complessità del modello) perché tende a minimizzare la perdita di informazione e scegliere modelli con migliori performance predittive, a costo di essere più complessi rispetto al BIC.

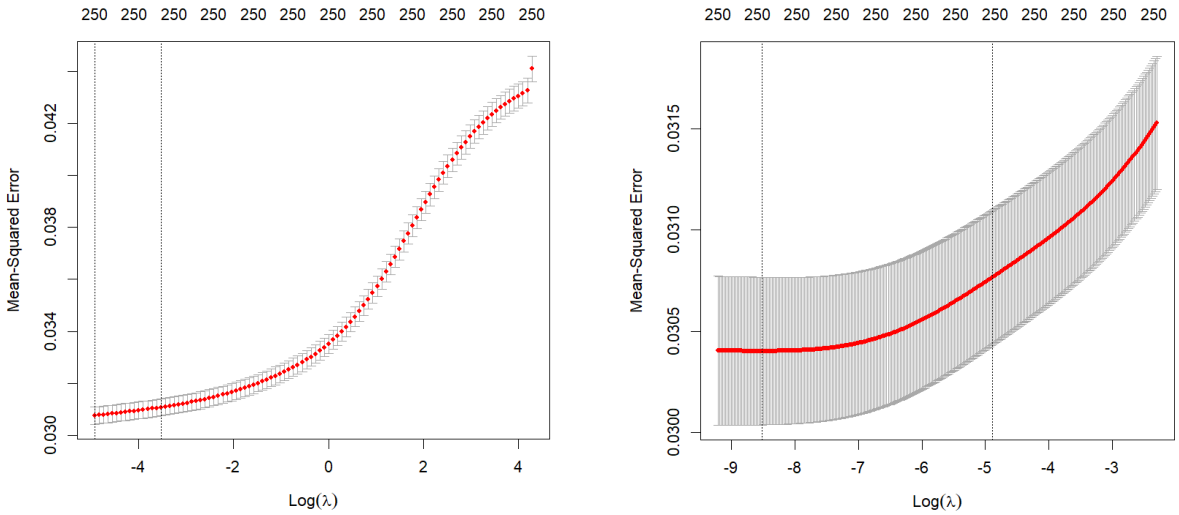
6. Modelli

Lo scopo dell'analisi è capire l'influenza delle variabili sulla risposta; dunque, privilegerò la scelta di modelli interpretabili e che tengono conto dell'alta correlazione tra le variabili, aiutandomi a gestirla.

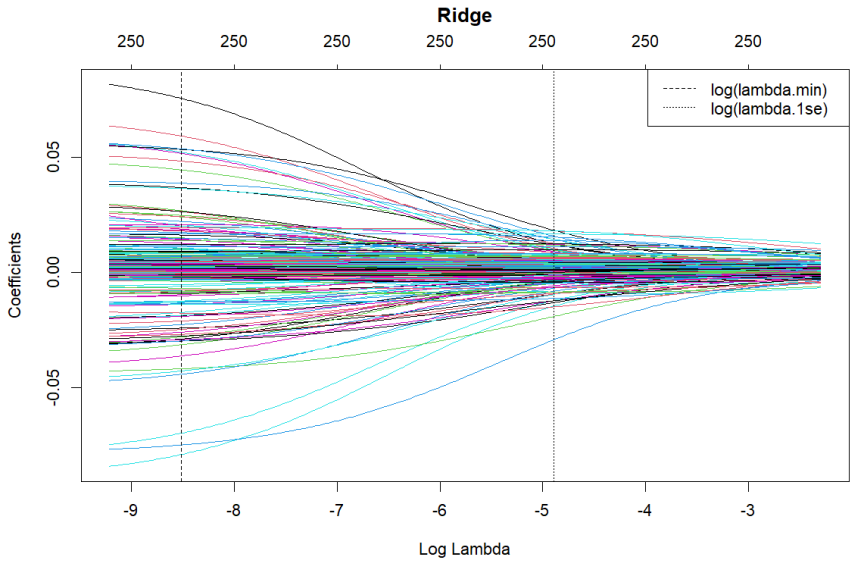
6.1 Modello Ridge

Inizio con una regressione lineare una regolarizzazione di norma L2 nella speranza di ridurre la multicollinearità e la varianza dei coefficienti stimati ma cercando di non introdurre troppa distorsione. Per farlo necessito di un parametro di regolazione lambda continuo positivo, anch'esso da stimare tramite la validazione incrociata dei dati divisi in 5 sottogruppi.

Una prima stima usa la griglia automatica di R per avere un'indicazione su dove concentrare la ricerca di lambda (in questo caso 100 valori compresi tra 0,007 e 72,490). Il valore minimo dell'errore al variare di log(lambda) è appena visibile nel primo grafico, dunque esploro altri 300 valori diversi compresi tra 10^-4 e 10^2 ed ottengo un lambda minimo pari a 0.00019 (secondo grafico).

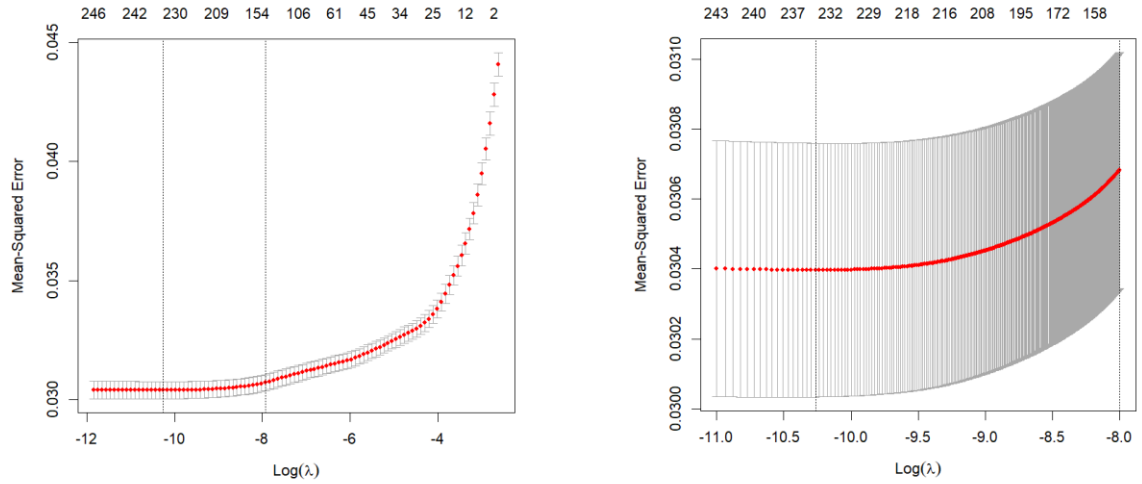


Di seguito il profilo dei coefficienti al variare di log(lambda) in corrispondenza del lambda minimi e del lambda 1se.

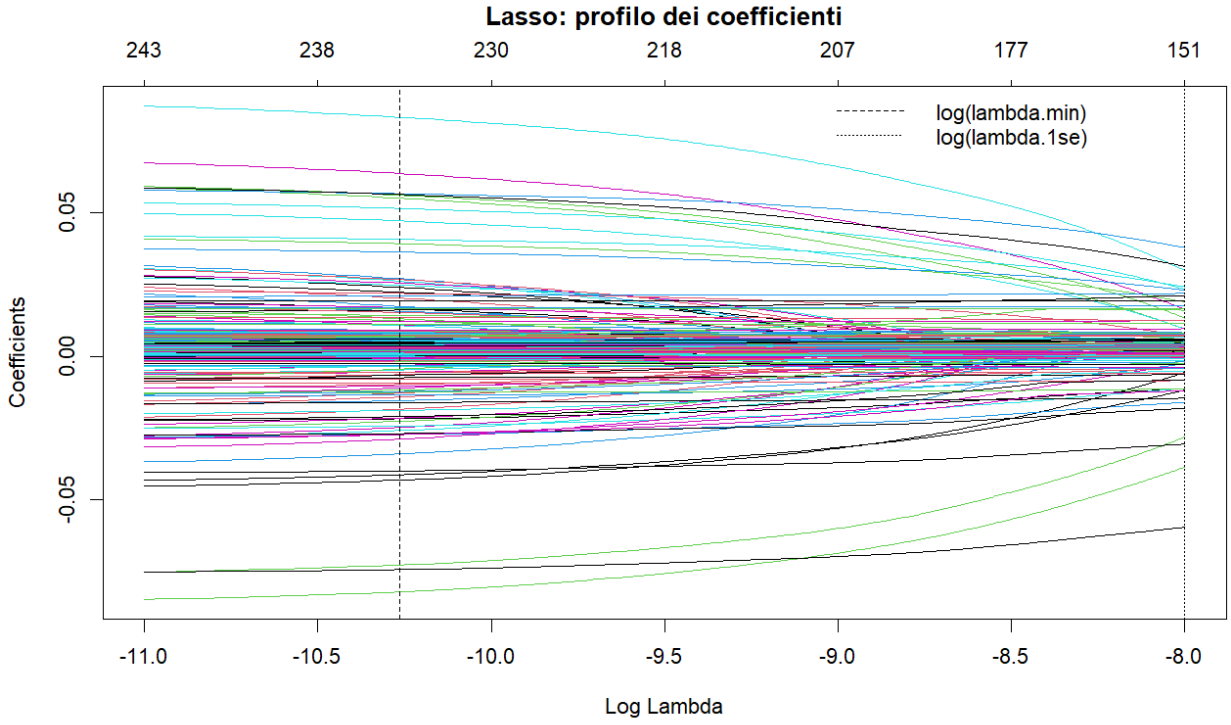


6.2 Modello Lasso

Come sopra, stimo il modello Lasso con la convalida incrociata con una griglia di 100 valori di lambda compresa tra 7,244*10^-6 e 7,24*10^-2, e 5 sottogruppi. Dal grafico a sinistra di rappresentazione dell'MSE al variare di log(lambda), mi accorgo che funzionano meglio valori piccoli di lambda, suggerendo una soluzione di ridotta selezione dei coefficienti rispetto al modello lineare. Seleziono il parametro di regolazione in una nuova griglia di 300 valori compresi tra e^-11 e e^-8, ottenendo un lambda minimo di 3,48e-05 (grafico a destra).



I coefficienti posti a zero sono 236 su 250. Segue il grafico del profilo dei coefficienti in corrispondenza del lambda minimo e del lambda 1se.



6.3 Modello Additivo con variabili selezionate dal lasso

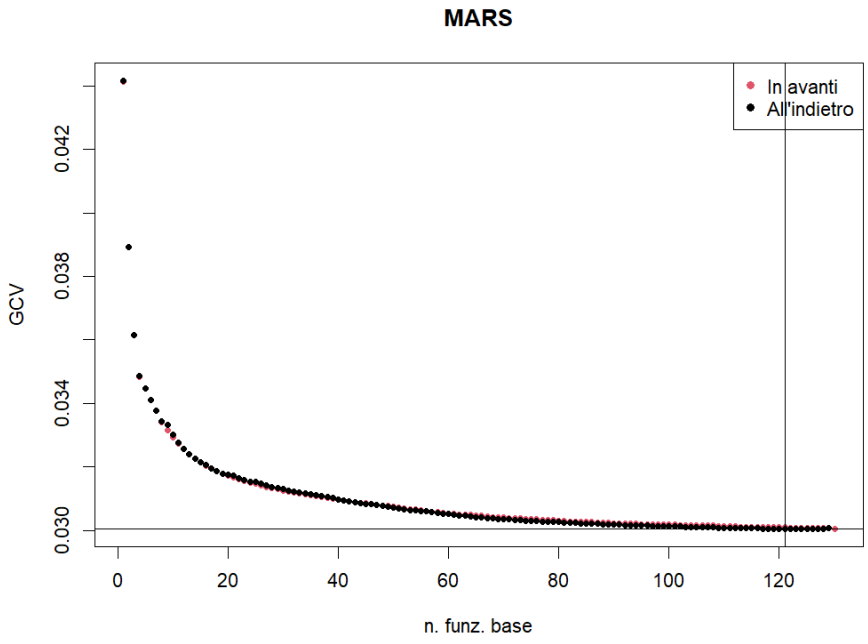
Lo calcolo sulle sole variabili selezionate dal lasso, senza includere interazioni. Uso spline cubiche di lisciamento (grado 3) per tutte le variabili quantitative (posso farlo perché ogni quantitativa ha almeno un numero di valori osservati pari al grado delle spline).

6.4 MARS

Stimo il MARS con la matrice del modello come predittori, basi additive e senza interazioni.

Imposto inizialmente un numero di funzioni di base massimo pari a 100, e poi, vedendo che non viene raggiunto un punto di minimo, lo aumento a 130. Il numero di nodi è fissato pari a 20. Il criterio di arresto è impostato con una tolleranza di 0,001.

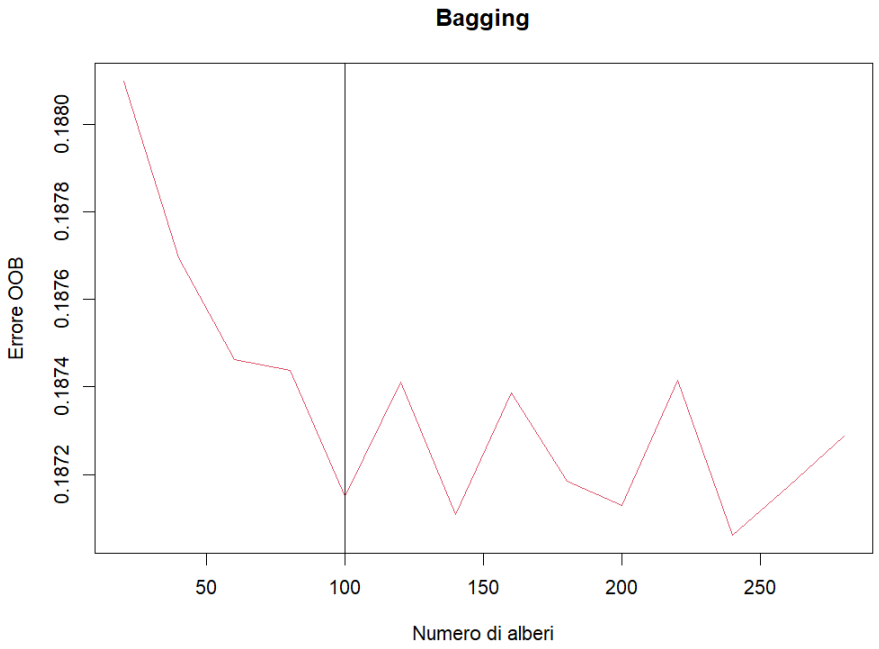
Il seguente grafico mostra l'andamento della stima dell'errore di previsione penalizzato per la complessità (GVC) al variare del numero di funzioni base (complessità) durante le fasi di crescita e potatura con una linea orizzontale in corrispondenza del minimo ed una verticale in corrispondenza del numero ottimo di funzioni base. Il valore scelto è 121.



6.5 Bagging

Trasformo la variabile COUNTRY_NAME con 61 modalità in 61 dummies corrispondenti a ciascuna modalità (lo stato di appartenenza del rispondente).

Stimo vari modelli al variare del numero di alberi con valori compresi tra 20 e 280 con passo 20, da cui ricavo il valore del parametro di regolazione in corrispondenza del punto in cui si stabilizza l'errore. Uso l'errore Out Of Bag, calcolato sempre per mezzo della split validation. Il seguente grafico mostra l'errore al variare del numero di alberi, con una retta verticale in corrispondenza del valore scelto di stabilizzazione.

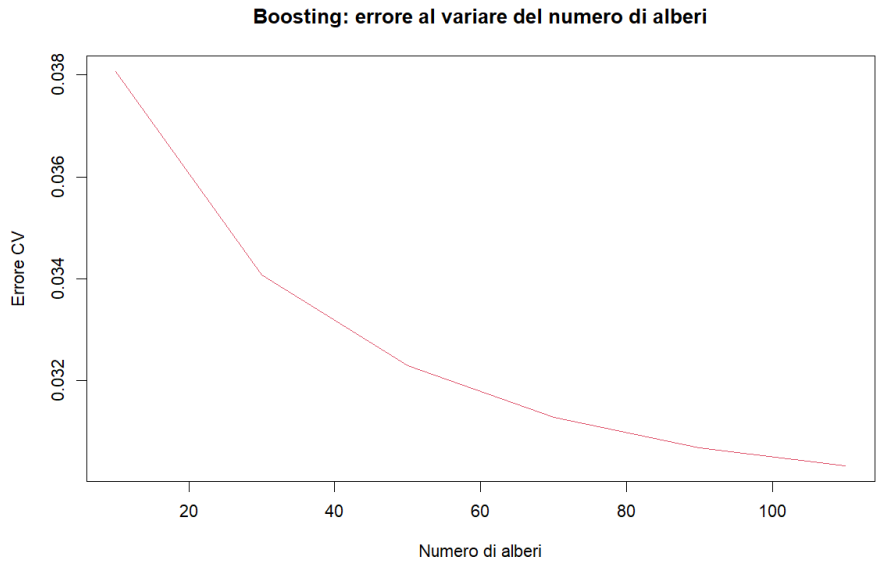


Calcolo il modello finale utilizzando un numero di alberi pari a 100.

6.7 Boosting

Seleziono il modello tramite convalida incrociata facendo variare il numero di alberi (da 10 a 120 con passo 20), impostando la profondità massima degli alberi a 3, una frazione di dati out of bag pari a 0,8 e 5 fold di convalida.

Ricavo dapprima il numero di iterazioni su tutti i dati di stima ed osservo che l'errore, nel range di iterazioni, non si stabilizza, ma diminuisce di pendenza in corrispondenza del valore 90. Ristimo, dunque, il modello finale utilizzando tutti i dati di stima ed un numero di alberi pari a 90.



7. Conclusioni

Di seguito, le tabelle con gli errori in ordine crescente ottenuti dai diversi modelli.

Modello	**MSE**
Additivo ridotto	0.03019
MARS	0.03062
Lineare Lasso	0.03070
Lineare Ridge	0.03070
Boosting	0.03215
Bagging	0.03594

Modello	**Log-lik**
MARS	0.5786
Lineare Lasso	0.5804
Lineare Ridge	0.5805
Boosting	0.5810
Additivo ridotto	0.5816
Bagging	0.5904

Osservo che il ranking dei modelli è identico, ad eccezione del Modello Additivo con variabili selezionate dal Lasso che si presenta al primo posto per quanto riguarda l'MSE e al quinto per quanto riguarda la perdita logaritmica.

La metrica più idonea per questo tipo di risposta (valori compresi tra 0 e 1) è la perdita logaritmica in quanto penalizza fortemente le previsioni troppo sicure (tendenti a 0 o 1) che sono errate, e moderatamente quelle incerte, garantendo una valutazione più informativa rispetto al semplice errore quadratico medio. Il modello indicato da questa metrica come migliore è il MARS, ma poiché l'obiettivo principale è interpretare le variabili, scelgo il modello Lasso di più facile interpretazione e leggibilità dei coefficienti.

Ordinando i coefficienti del Lasso per valori di beta crescenti (la standardizzazione permette questo confronto), ottengo che quelli nelle prime posizioni riguardano tutti modalità relative al paese di appartenenza COUNTRY_NAME, individuando tale variabile come la più utile ai fini della previsione. Di seguito vengono riportati i primi 10 coefficienti, potendo osservare che si tratta sempre di categorie di COUNTRY_NAME codificate come dummy.

COUNTRY_NAMECongo DR 8.349555e-02

COUNTRY_NAMECzech Republic -8.223503e-02

COUNTRY_NAMERussia -7.445582e-02

COUNTRY_NAMEBolivia -7.285356e-02

COUNTRY_NAMECameroon 6.383715e-02

COUNTRY_NAMEIsrael 5.677502e-02

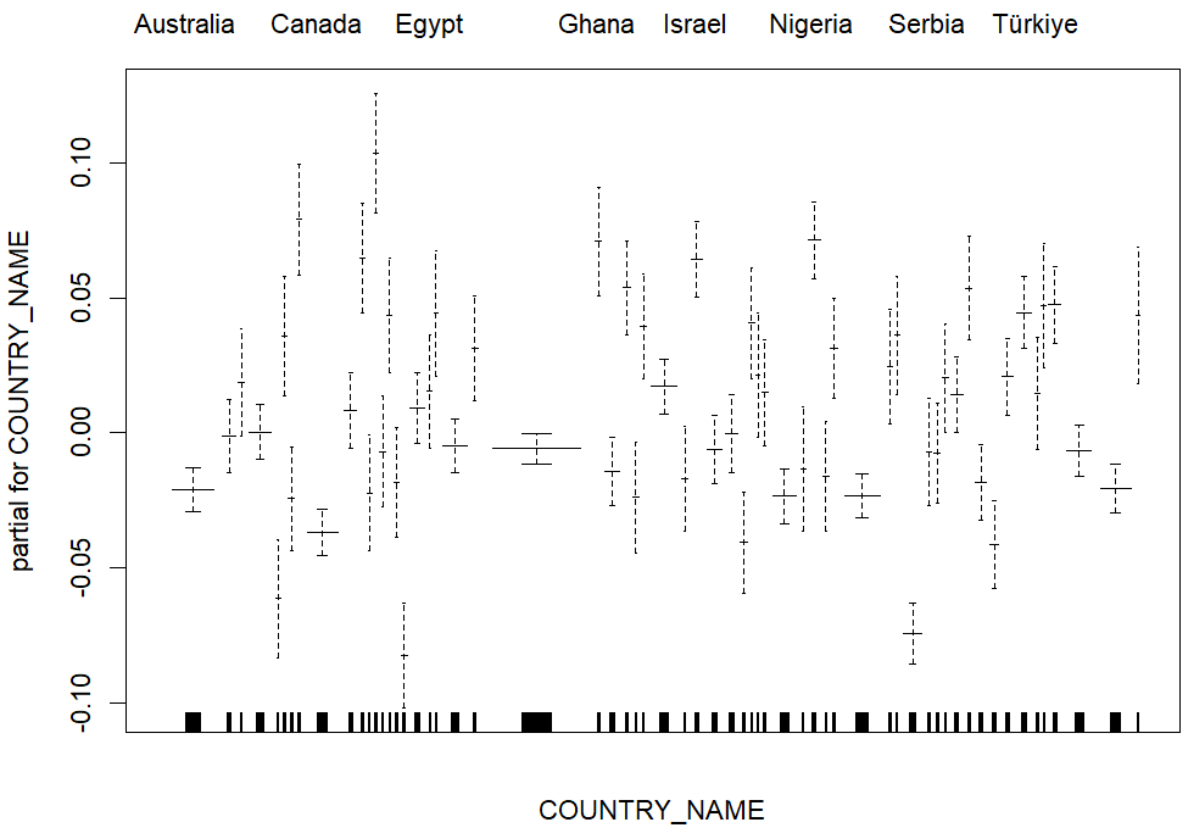
COUNTRY_NAMENigeria 5.650466e-02

COUNTRY_NAMEChina 5.631309e-02

COUNTRY_NAMEGhana 5.520920e-02

COUNTRY_NAMEHong Kong 5.162392e-02

Per quanto riguarda l'MSE, il modello migliore è quello Additivo dove è possibile osservare come varia il contributo individuale della variabile COUNTRY_NAME al variare delle modalità tramite il suo partial plot riportato qui di seguito (per leggerlo, i Paesi sono ordinati in ordine alfabetico).



In conclusione, il Paese di appartenenza emerge come il principale fattore che spiega il livello di supporto dei cittadini alle politiche di contrasto al cambiamento climatico.