Michela Ropele

Data Mining: Energia

1. Quesito di ricerca

Un'azienda fornitrice di energia elettrica che opera nell'intero territorio italiano è interessata ad analizzare alcuni dati relativi al consumo di energia nella regione Lombardia. La previsione dei consumi della propria customer base garantisce un vantaggio competitivo alle aziende che lavorano nel mercato dell'energia elettrica. L'azienda è, quindi, interessata a prevedere i consumi di energia elettrica dei propri clienti nei mesi futuri, ma anche a stimare quale potrebbe essere il consumo di potenziali nuovi clienti, anche residenti in aree geografiche diverse, seppure vicine, a quelle dove risiedono gli attuali clienti.

Per affrontare queste tematiche l'azienda vi mette a disposizione un dataset (energia.csv) contenente una serie di informazioni relative ai contratti di energia elettrica relativi alla customer base lombarda. Le variabili disponibili sono riportate e descritte in appendice.

2. Dati

Il dataset presenta inizialmente 187172 osservazioni e 29 variabili, e non sono presenti osservazioni ripetute. Le unità statistiche sono i singoli punti di fornitura dei vari clienti in un certo mese. La variabile risposta è il consumo di energia mensile in kWh, quantitativa continua che assume valori reali positivi.

L'obiettivo è prevedere il consumo energetico futuro dei clienti attivi e stimare quello dei nuovi clienti (obiettivo predittivo). Si intende valutare anche la presenza di eventuali dipendenze spaziali e temporali nei consumi.

Eseguo le seguenti operazioni di pulizia:

- Rimuovo l'identificativo di riga;
- Rimuovo la variabile "kWh_giorni" perchè ottenuta dalla risposta divisa per il numero di giorni mensili;
- Rimuovo le variabili testuali relative al comune ("Comune_sede", "Comune_sed", "comune_fornitura" e "comune"), mantenendo invece le coordinate geografiche ("latitudine", "longitudine") e la superficie territoriale ("superficie"), che forniscono informazioni quantitative più precise;
- Faccio l'analogo con le variabili relative all'altitudine: tengo "altitudine" e rimuovo "montano" e "zona-altimetrica";
- Rimuovo "mese_nascita_cliente" e "giorno_nascita_cliente";
- Aggiungo una categoria "non persona" alla variabile che indica il sesso del cliente, per distinguere i casi in cui il cliente è un'azienda o un ente (per i quali il sesso non è applicabile);
- Sostituisco i valori anomali 0 e 1 con 100 e 101, in modo da interpretarli correttamente come anni di nascita 2000 e 2001. Successivamente, la trasformo in categoriale per distinguere le classi di anno di nascita dei clienti privati, dall'assenza di questa informazione per i clienti diversi dai privati;
- Rimuovo anche l'identificativo dell'edificio "Id_Sito" per concentrarmi solo sull' identificativo del cliente;

Ora il dataset presenta 187172 osservazioni e 18 variabili.

3. Operazioni preliminari

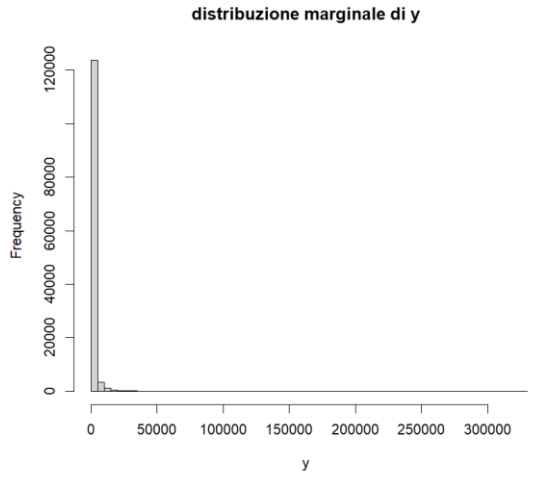
Il dataset viene suddiviso in due insiemi, stima (70%) e verifica (30%), assicurandosi che la suddivisione avvenga a livello di cliente ("Id_Cliente") per evitare che lo stesso cliente compaia in entrambi i set. Si suddivide ulteriormente l'insieme di stima in stima e convalida per la selezione dei parametri di regolazione (2/3 e 1/3 dei clienti). Si imposta un seme casuale per garantire la riproducibilità dei risultati in tutte le operazioni che richiedono la generazione di numeri casuali.

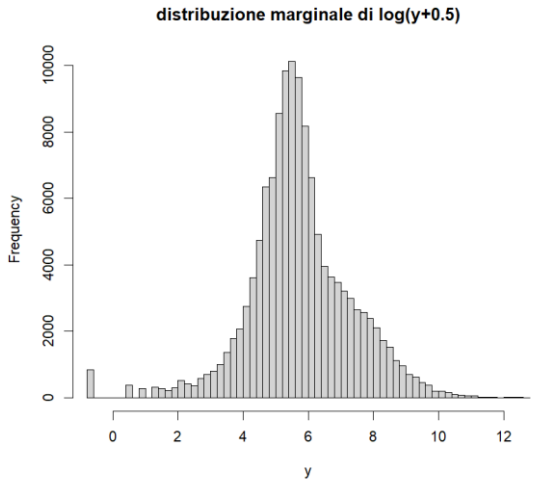
Le variabili quantitative vengono standardizzate separatamente nei set di stima e verifica, in modo da rendere i coefficienti confrontabili e prevenire che la scala influenzi modelli sensibili alla magnitudine delle variabili.

Le variabili "consumo_annuo" e "Potenza" hanno correlazione pari a 0.74 e ciò potrebbe creare dei problemi di collinearità nei modelli lineari. Tuttavia, scelgo di non rimuoverle preventivamente in modo da lasciare questa selezione ai modelli con penalizzazione o selezione automatica.

4. Variabile risposta

La distribuzione marginale della variabile d'interesse mette in luce valori sempre positivi e con un range di variazione di diversi ordini di grandezza, a causa dei quali la distribuzione della risposta risulta fortemente asimmetrica. Applico la trasformazione log(y + 0.5) per ridurre l'asimmetria, stabilizzare la varianza ed evidenziare una relazione più lineare con i predittori. Questa trasformazione consente un'interpretazione dei coefficienti in termini di variazioni percentuali più adatta al contesto economico dell'analisi.





5. Prestazioni del modello

Per valutare le diverse previsioni, utilizzo la metrica MSE che penalizza molto gli errori grandi e poco quelli piccoli. Se dovrò effettuare la scelta di un criterio di informazione utilizzerò l'AIC (ovvero la verosimiglianza penalizzata dalla complessità del modello) perché tende a scegliere modelli con buone capacità predittive, accettando un livello di complessità maggiore rispetto al BIC.

6. Modelli

6.1 Modello lineare stepwise

Includo l'interazione tra latitudine e longitudine e avvio la selezione stepwise partendo dal modello nullo, così da costruire progressivamente modelli più complessi solo se migliorano la qualità della stima. Inoltre, questa procedura seleziona in autonomia le variabili da inserire tra quelle ridondanti.

Il modello con regressione ibrida mi restituisce tutte le variabili, inclusa l'interazione lineare tra latitudine e longitudine.

Metrica: 1,2295

6.2 Modello GAM

Passo a calcolare un modello additivo con lisciatori univariati di grado 3 ed uno bivariato per l'interazione tra latitudine e longitudine. Inserisco le variabili selezionate dalla regressione lineare stepwise (tutte). L'interazione tra latitudine e longitudine è significativa, così come lo è il tempo.

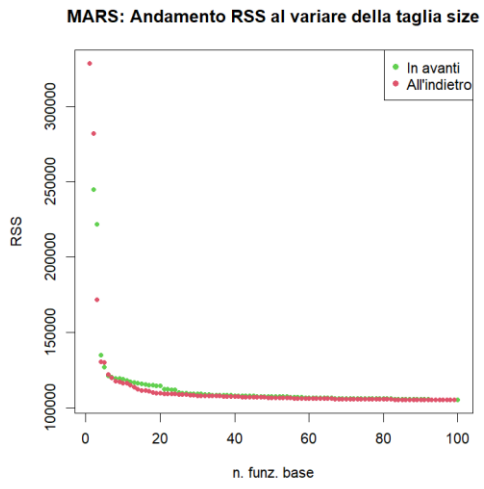
Metrica: 1,2049

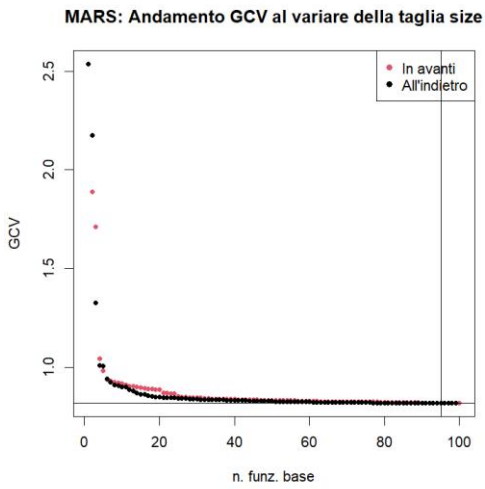
6.3 MARS

Uso il numero di funzioni di base massimo pari a 100 ed il numero di nodi calcolato di default pari a 20. Inserisco nel modello sia gli effetti principali che le interazioni (parametro additive = FALSE). Il criterio di arresto è impostato con una tolleranza di 0,001.

Nella crescita del modello uso la somma dei quadrati dei residui (RSS), mentre nella potatura la convalida incrociata generalizzata (GVC).

I seguenti grafici mostrano l'andamento della stima dell'errore di previsione penalizzato per la complessità (RSS e GVC) al variare del numero di funzioni base (complessità) durante le fasi di crescita e potatura con (nel secondo grafico) una linea orizzontale in corrispondenza del minimo ed una verticale in corrispondenza del numero ottimo di funzioni base. Il valore scelto è 93.





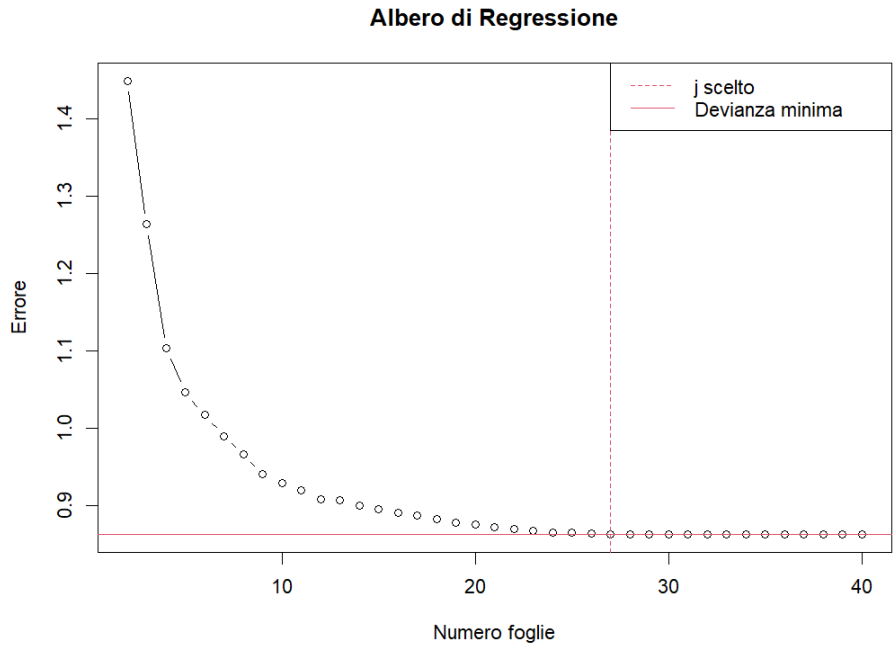
Metrica: 5,0689

Da ora in poi, le interazioni tra latitudine e longitudine saranno calcolate e modellate in modo implicito dai modelli.

6.4 Albero di regressione

Uso la convalida incrociata con 5 sottogruppi. Imposto l'arresto dell'algoritmo ad un numero minimo di 15 osservazioni per foglia e una devianza minima di 0,01. Il test viene eseguito su alberi con un numero di foglie j variabile da 2 a 40. Le interazioni tra latitudine e longitudine vengono modellate implicitamente.

Dal grafico di andamento della devianza media in funzione della taglia dell'albero, noto che raggiungo il minimo dell'errore medio di previsione dei diversi sottogruppi in corrispondenza di j=26, quindi 27 foglie.



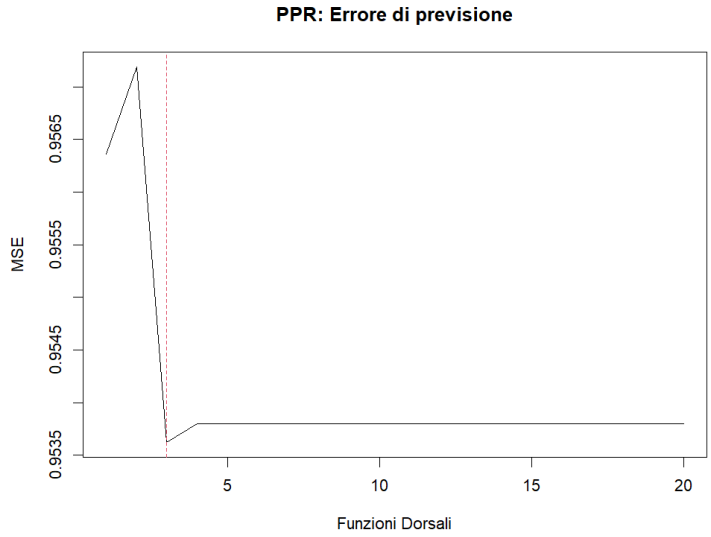
Stimo l'albero finale potato in corrispondenza di j sull'insieme con tutti i valori di stima. Non riporto il grafico dell'albero finale perché non è leggibile.

Metrica: 1,5739

6.5 Regressione Projection Pursuit (PPR)

Anche questo modello stima implicitamente interazioni non lineari attraverso le proiezioni delle variabili. Utilizzo un numero di funzioni dorsali (parametro di regolazione) variabile da 1 a 20.

Il grafico dell'errore di previsione stimato al variare delle funzioni dorsali è il seguente, con una retta tratteggiata in corrispondenza del parametro di regolazione con l'errore minimo.



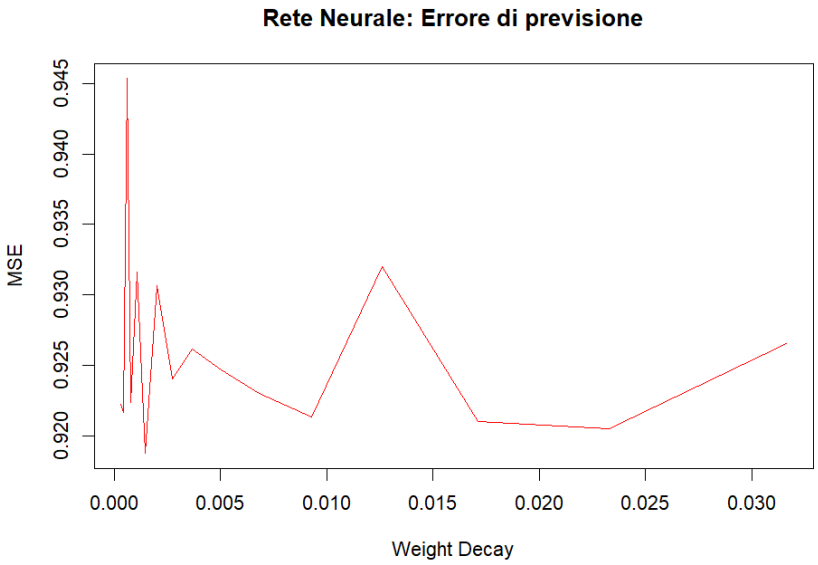
Il valore scelto è 3. Ricalcolo il modello con 3 funzioni dorsali.

Metrica: 5,5438

6.6 Rete Neurale

Considero le sole variabili quantitative, tenendo presente che il modello restituito non sarà interpretabile, ma potrebbero comunque modellarsi interazioni implicite tra le variabili spaziali.

Imposto una griglia standard di weight decay in scala logaritmica di 16 valori compresi tra 10^-3.5 e 10^-1.5. Fisso già 5 nodi nello strato nascosto e scelgo l'output lineare idoneo per eseguire regressione. Imposto il numero di iterazioni massime a 700.



Identificato il peso di decadimento ottimale pari a 0.00147, stimo il modello su tutto l'insieme di stima.

Metrica: 55,016

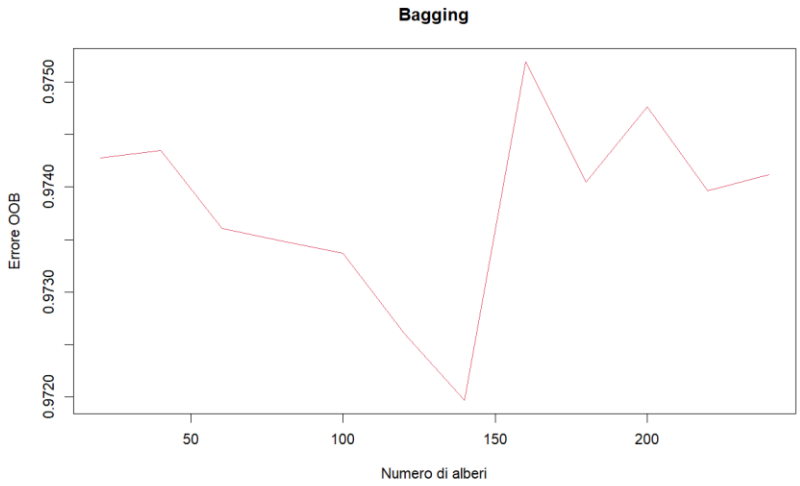
L'errore qui è particolarmente alto, probabilmente per colpa del fatto che sono state incluse le sole variabili quantitative e il numero di unità nello strato latente è basso. Ricalcolo il modello aumentando questo valore a 40 e diminuendo il numero di iterazioni massime a 200 per mantenere i tempi di calcolo contenuti.

Nuova metrica: 10.4623, ma è ancora superiore alle precedenti.

Il modello potrebbe essere ulteriormente migliorato aumentando il numero di nodi e facendo una nuova valutazione anche per il weight decay.

6.7 Bagging

Effettuo la riduzione del dataset e stimo modelli al variare di un numero di alberi che va da 20 a 240 a passo 20, da cui ricaverò il valore del parametro di regolazione ottimale in funzione di dove si stabilizza l'errore. Uso l'errore Out Of Bag.



Non si osserva una chiara stabilizzazione dell'errore Out-Of-Bag, ma scelgo 160 alberi come compromesso tra complessità e accuratezza. Infatti, la metrica è contenuta ed inferiore a quella dell'albero.

Metrica: 1,5992

7. Conclusioni

Il modello additivo generalizzato (GAM) risulta quello con le migliori prestazioni predittive, seguito dal modello lineare stepwise. Di seguito la tabella degli errori (MSE) per i vari modelli.

Modello	**MSE**
GAM	1.2049
Lineare passo passo	1.2295
Albero di regressione	1.5739
Bagging	1.5992
MARS	5.0689
Regressione Projection Pursuit	5.5438
Rete neurale	10.4623

L’output del modello additivo GAM mostra che sia le variabili spaziali che quelle temporali contribuiscono in modo significativo alla previsione della risposta, come evidenziato dai valori di significatività approssimativa dei termini:

Approximate significance of smooth terms**:**

	edf	**Ref.df**	**F**	**p-value**
s(long)	1.922	1.989	13.795	4.25e-06 ***
s(lati)	1.743	1.919	1.823	0.2274
s(time)	1.000	1.000	4.763	0.0291 *
s(lati,long)	26.901	27.000	29.841	< 2e-16 ***
s(altitudine)	1.976	1.999	33.656	< 2e-16 ***
s(superficie)	1.974	1.999	151.710	< 2e-16 ***

Anche il modello di regressione lineare conferma la significatività di queste variabili, inclusa l’interazione tra latitudine e longitudine:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
long	0.010877	0.003505	3.104	0.00191 **
lati	-0.051850	0.003633	-14.272	< 2e-16 ***
time	-0.010876	0.003565	-3.051	0.00228 **
lati:long	-0.010980	0.002697	-4.071	4.69e-05 ***
superficie	-0.046030	0.003483	-13.217	< 2e-16 ***
altitudine	-0.027328	0.003741	-7.304	2.81e-13 ***

Per la previsione dei clienti attuali, il modello selezionato può essere applicato utilizzando tutte le informazioni disponibili. Per stimare i consumi dei nuovi clienti, è possibile utilizzare lo stesso modello, impostando la variabile “time” ai valori corrispondenti al periodo di previsione, poiché non si dispone di dati storici precedenti.

Appendice: variabili

Id_Cliente Identificativo di cliente

Id_Sito Identificativo del punto di fornitura

Categoria_cliente Tipologia di cliente:

AZN: azienda

PRIV: privato

PAMM: pubblica amministrazione

COND: utenza condominiale

NPRO: azienda noprofit

GENP: generico pubblico (altro)

anno_nascita_cliente Anno di nascita del cliente (per clienti privati e partite iva)

mese_nascita_cliente Mese di nascita del cliente (per clienti privati e partite iva)

giorno_nascita_cliente Giorno di nascita del cliente (per clienti privati e partite iva)

sesso_cliente Genere del cliente (per clienti privati e partite iva)

0: femmina 1: maschio

Comune_sede Codice Istat del Comune sede del cliente

Comune_sped Codice Istat del Comune dove viene inviata la fattura

Fattura_Via_Email Richiesta del cliente di ricevere la fattura via e-mail

1: Sì 0: No

Comune_fornitura Codice Istat del comune in cui è installato il punto di fornitura

Consumo_Annuo Consumo annuo in kWh dichiarato dal cliente

in fase di sottoscrizione del contratto

Opzione_tariffaria Opzione tariffaria (variabile qualitativa con 13 modalità)

Potenza Potenza disponibile in kWh per punto di fornitura

Tipo_pagamento Tipologia pagamento della fattura:

Addebito SDD

bollettino postale

bonifico

Scadenza Tipologia Scadenza (variabile qualitativa con 2 modalità)

comune Denominazione in italiano del Comune sede del cliente

zona_altimetrica Zona altimetrica del comune

1: montagna interna

3: collina interna

5: pianura

altitudine Altitudine s.l.m in metri in corrispondenza della sede del Municipio del Comune.

montano Variabile qualitativa

1: Non montano

2: Parzialmente montano

3: Totalmente montano

superficie Superficie territoriale del comune in kmq

urbanizzazione Grado di urbanizzazione

1: densamente popolato

2: densità intermedia

3: scarsamente popolato (rurale)

long Longitudine del Municipio del Comune

lati Latitudine del Municipio del Comune

time Indicatore sequenziale del mese di osservazione

1: ottobre 2016 ... 36: settembre 2019

mese Indicatore del mese dell'anno

1: gennaio 2: febbraio ... 12: dicembre

kWh Consumo mensile di energia elettrica in kWh

kWh_giorn Consumo medio giornaliero per mese di energia elettrica in kWh

per ciascun mese si ottiene la media giornaliera dei consumi attraverso la divisione del consumo mensile per il numero di giorni