LUISS

Corso di laurea in Economia e Management

Cattedra di Informatica

Previsione del prezzo dell'energia elettrica in Italia: un'analisi comparativa di modelli di machine learning

Prof. Fabio Dellutri
RELATORE

Lorenzo Massari Matr. 266361
CANDIDATO

Anno Accademico 2023/2024

ABSTRACT

La deregolamentazione dei mercati dell'energia e le crisi globali recenti, come il COVID-19 e l'invasione dell'Ucraina da parte della Russia, hanno posto nuove sfide e opportunità per la previsione dei prezzi dell'elettricità, un elemento importante della pianificazione strategica del settore energetico. Lo scopo di questa tesi è quello di predire il Prezzo Unico Nazionale (PUN) del'energia elettrica in Italia utilizzando il machine learning. Saranno confrontati vari modelli, tra cui: Support Vector Regression, Random Forest, Gradient Boosting, XGBoost e regressione lineare. Verranno considerate come input diverse variabili, tra le quali: i prezzi delle materie prime, il rischio geopolitico, la domanda di energia ed i dati climatici. I risultati mostrano che il gas naturale influenza in modo più significativo il PUN, con i modelli SVR e regressione Ridge che eccellono per le loro capacità predittive. In più, un approccio di ensemble basato sui modelli più performanti ha ulteriormente migliorato la prestazione del modello in termini di accuratezza e stabilità. In questo caso, il modello risultante può essere utilizzato ad esempio per massimizzare la produzione energetica o per capire quando negoziare i contratti, incrementando così gli strumenti utili agli operatori del mercato.

INDICE

1 INTRODUZIONE

Negli ultimi decenni il settore dell'energia è stato rivoluzionato, grazie ad una massiccia liberalizzazione dei mercati energetici. Questo ha permesso un alto livello di concorrenza, facendo sì che fosse necessario per i player del settore avere modelli sofisticati per la previsione di prezzi e consumi. La predizione di questi due elementi è diventata una parte integrante dei business e delle decisioni di molteplici attori, dai produttori ai gestori, dai venditori all'ingrosso fino ai consumatori finali.

Il mercato elettrico italiano è conosciuto come IPEX, ovvero Italian Power Exchange, ed è uno dei più interessanti case study del settore per la trasparenza e la facilità di ottenimento di dettagliati dati al riguardo. Il mercato italiano è organizzato in zone e basato su un Prezzo Unico Nazionale (PUN) che è una media ponderata dei vari PUN zonali. Questo sistema consente una valutazione dettagliata delle criticità e dei benefici associati alla continua integrazione dei mercati (obiettivo auto-proclamato dell'Unione Europea).

Secondo la letteratura, nel settore energetico, le previsioni si basavano sulla teoria dei modelli autoregressivi lineari; tali modelli erano noti per la loro semplice implementazione. Adesso, le nuove tecniche di apprendimento automatico riescono ad apportare diversi vantaggi, in quanto consentono di integrare un maggiore numero di variabili esplicative e di catturare meglio la complessità dei dati. Tuttavia, gli studi nella letteratura sopracitata constatano che l'efficacia di questi nuovi metodi varia a seconda del contesto e del modello.

Questa tesi ha due obiettivi principali: innanzitutto, cercare di sviluppare previsioni accurate dei prezzi dell'energia elettrica utilizzando algoritmi di machine learning e, in secondo luogo, studiare qual è l'influenza delle variabili su tutte le altre analizzate. Verranno creati diversi modelli che tengano conto della sequenza temporale, risultando in un modello effettivamente utile ed utilizzabile per la pianificazione e la gestione del mercato energetico.

La tesi comprende tre capitoli principali: Il secondo capitolo presenta l'evoluzione del mercato italiano. Esso illustra lo sviluppo del mercato energetico ed il suo funzionamento. Il terzo capitolo include le metodologie di previsione adottate, incluse sia tecniche statistiche tradizionali che modelli di machine learning avanzati. Infine, i risultati sono

riassunti nel quarto capitolo. Tale capitolo pone a confronto le prestazioni dei diversi modelli di apprendimento automatico.

2 LAVORI SIMILI

L'argomento della previsione dei prezzi e della domanda di elettricità è ampliamente trattato in letteratura, soprattutto dopo la liberalizzazione dei mercati energetici. Tra i lavori che hanno influenzato la stesura di questa tesi spiccano due tesi dell'Università di Padova e una dell'Università di Stoccolma [23,24,25], che si focalizzano sul processo di modellazione dei prezzi utilizzando machine learning e modelli autoregressivi. Tuttavia, nonostante le tesi dell'Università di Padova riguardino le dinamiche del mercato elettrico italiano e i modelli di previsione, come le Support Vector Machines, Random Forest e Gradient Boosting, queste basano i loro risulti su dati fino al 2021.

Mentre i lavori sopra citati offrono una base solida per capire le dinamiche dei prezzi dell'energia all'interno del sistema nazionale, essi non toccano le problematiche di mercato significative che sono sorte a seguito della crisi energetica esplosa a livello Europeo nel 2021-2022. Questa tesi differisce includendo un'analisi delle variazioni di prezzo più recenti, consentendo un insight più pertinente per il mercato attuale. Ciò è di grande importanza perché il contesto della crisi energetica ha fornito nuovi driver di prezzo, come l'aumento dei prezzi del gas naturale e le conseguenze della guerra in corso in Ucraina, che non sono stati considerati da tesi precedenti.

Per quanto riguarda invece la tesi dell'Università di Stoccolma, questa ha un approccio più innovativo, implementando modelli di deep learning, come LSTM e TCN, alla previsione dei prezzi dell'elettricità del mercato dei paesi nordici. Questa tesi è stata fondamentale per suddividere i capitoli, organizzare e strutturare l'analisi metodologica della seguente tesi. Tuttavia, mentre il progetto dell'università di Stoccolma si focalizza principalmente sul mercato dei paesi nord-europei e sui contratti futures, questa tesi è incentrata sul mercato italiano e include un'analisi comparativa tra diversi modelli di machine learning e la creazione di un modello ibrido basato su variabili esogene.

3 CONTESTO

3.1 Come funziona il mercato energetico italiano e cos'è il PUN

Il mercato energetico italiano è un sistema complesso che è stato creato al fine di fornire energia elettrica a tutti gli utenti in modo quanto più efficiente e sicuro possibile. Il processo di liberalizzazione del mercato energetico ha avuto inizio con l'emanazione del Decreto Legislativo n. 79/1999, noto come Decreto Bersani, il quale ha recepito la Direttiva europea 96/92/CE. L'obiettivo del decreto è stato quello di passare da un sistema monopolistico, con Enel come unico fornitore, a un sistema aperto alla concorrenza, in modo da promuovere efficienza e competitività e trasparenza.

Il mercato energetico italiano si divide principalmente in due segmenti:

1) Mercato all'ingrosso: dove l'energia viene scambiata tra produttori e grossisti attraverso compravendita. Questo segmento comprende diversi mercati gestiti dal Gestore dei Mercati Energetici (GME) e si articola in:
    i. Mercato del Giorno Prima (MGP): qui si svolgono le contrattazioni per la fornitura di energia elettrica del giorno successivo.
    ii. Mercato Infragiornaliero (MI): consente ulteriori contrattazioni per permettere ai produttori di "modellare" i programmi di produzione ottenuti dal MGP.
    iii. Mercato dei prodotti giornalieri (MPEG): qui si negoziano prodotti giornalieri con obbligo di consegna dell'energia.
    iv. Mercato del servizio di dispacciamento (MSD): attivato in caso di congestioni o per l'impiego di riserve di elettricità.
    v. Mercato a Termine dell'energia (MTE): dove si negoziano contratti a lungo termine per la fornitura di energia

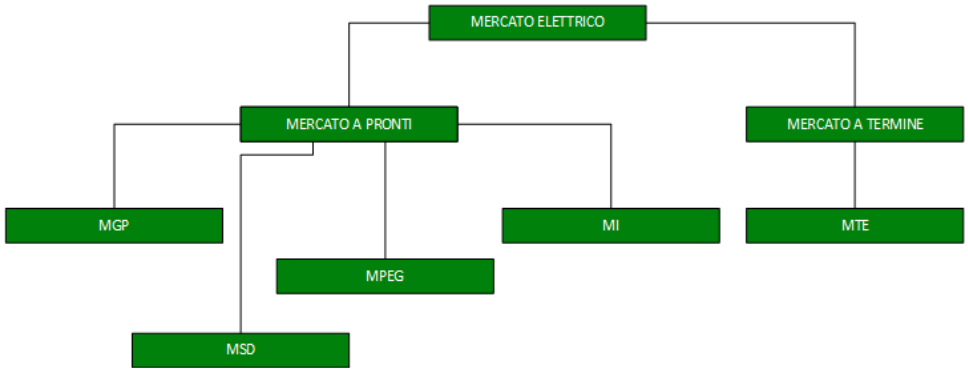

Figura 1: Come si articola il mercato elettrico, fonte: GME

2) Mercato al dettaglio: dove l'energia viene venduta ai consumatori finali. Qui i fornitori acquistano l'energia dal mercato all'ingrosso e la rivendono ai clienti finali tramite i contratti di fornitura.

Sulla base delle transazioni che avvengono nel MPG viene calcolato il Prezzo Unico Nazionale (PUN), elemento chiave nel mercato all'ingrosso in quanto rappresenta il prezzo di riferimento per l'energia elettrica. Si tratta quindi del prezzo dell'energia all'ingrosso espresso in €/MWh e viene calcolato utilizzando due valori: il prezzo offerto dai fornitori per l'acquisto di energia elettrica sul mercato e i prezzi dell'energia elettrica proposti nelle diverse zone d'Italia. Alla chiusura del MGP viene effettuata la media ponderata dei valori del PUN raccolti ad intervalli orari ottenendo così il valore del PUN del giorno.

Si evince facilmente quanto il ruolo del PUN sia cruciale per il mercato elettrico italiano: oltre ad essere uno degli indicatori di performance principali del mercato elettrico, è anche il principale segnale di prezzo che riflette l'equilibrio tra domanda e offerta di energia elettrica a livello nazionale. Difatti, con riferimento al mercato al dettaglio (ma anche quello all'ingrosso) è il prezzo che viene usato come riferimento in molti contratti di fornitura di energia elettrica.

3.2 Da cosa dipende generalmente il prezzo dell'energia

Molti sono i fattori che generalmente influenzano l'andamento del prezzo dell'energia a livello globale, tra questi sicuramente identifichiamo [13]:

1) Domanda e offerta: generalmente, così come accade nella vita di tutti i giorni, quando la domanda supera l'offerta il prezzo tende a salire. Viceversa, quando l'offerta supera la domanda allora i prezzi scendono. Allo stesso modo funziona il prezzo dell'energia elettrica. L'effetto di questo fenomeno può essere anche enfatizzato nel mercato elettrico considerando il fatto che è influenzato dalle stagionalità (riscaldamenti in inverno e aria condizionata in estate) oltre che dagli eventi straordinari come ondate di calore e di gelo.
2) Il costo delle materie prime: soprattutto per quanto riguarda l'energia termoelettrica, ovvero l'energia generata tramite combustione, il prezzo dei combustibili (come gas, petrolio, carbone, biocarburanti, ecc.) può influenzare

notevolmente il prezzo dell'energia elettrica. Un innalzamento repentino del prezzo per petrolio, ad esempio, può causare l'aumento del prezzo dell'elettricità.

3) Politiche governative e regolamentazioni: le decisioni politiche, anche comunitarie, come la tassazione sulle emissioni di CO_2, sussidi per le energie rinnovabili o normative ambientali possono influenzare i prezzi dell'energia, ad esempio, rendendo molto più dispendiosa l'energia generata dai combustibili fossili anziché quella da fonti rinnovabili.
4) Eventi geopolitici: le tensioni geopolitiche, come conflitti o sanzioni internazionali possono alterare l'equilibrio del mercato elettrico di un paese anche non direttamente coinvolto.
5) Tendenze macroeconomiche: fattori come il PIL, l'inflazione ed i tassi d'interesse anche se non direttamente possono modificare l'equilibrio tra domanda e offerta. In un'economia in crescita, ad esempio, la domanda energetica tende ad aumentare.
6) Clima e condizioni meteorologiche: in un paese con una buona quota di energia rinnovabile, come fotovoltaico ed eolico, il prezzo dell'energia può variare anche di molto in base a clima e meteo. Ad esempio, in estate, con temperature più elevate e giornate generalmente più soleggiate, la produzione fotovoltaica risulterà essere di gran lunga maggiore rispetto all'inverno.

3.3 Fonti di Generazione di Energia Elettrica in Italia.

L'Italia per soddisfare il suo fabbisogno energetico impiega diverse fonti sia rinnovabili che non.

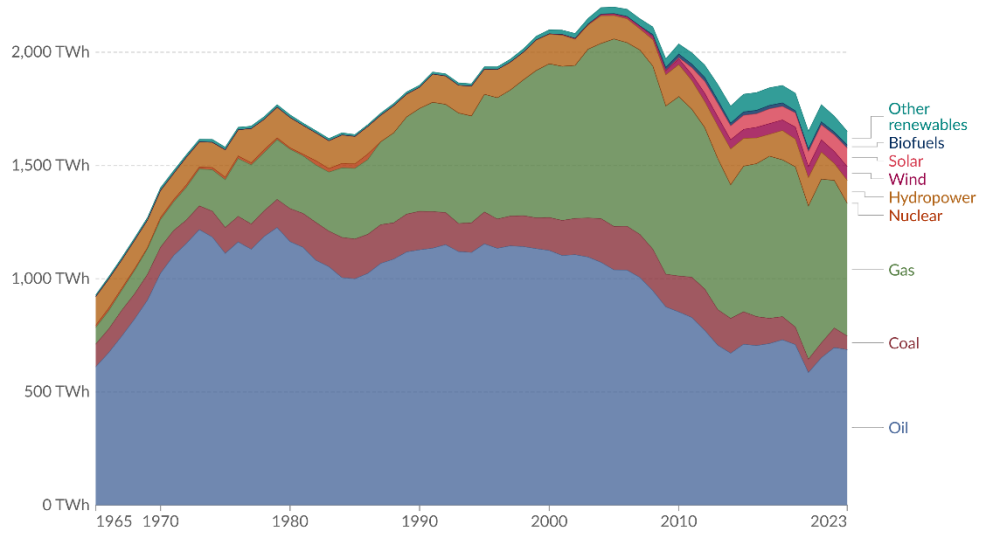

Figura 2: Fonti di generazione di energia elettrica in Italia, fonte: Our World in Data

Dal grafico riportato si evince subito come l'Italia basi la maggior parte della produzione su fonti non rinnovabili, come petrolio ("O*il*"), carbone ("*Coal*") e gas, quasi totalmente importati dall'estero.

La quota derivante da fonti rinnovabili, invece, pur essendo esigua, è in una lenta ma costante crescita grazie agli investimenti e alle regolamentazioni comunitarie: grazie al PNRR, nella sezione "*rivoluzione verde e transizione ecologica*" (con un ammontare di 59,52 miliardi di euro)[1] sono stati presentati ed avviati progetti di potenziamento di queste infrastrutture.

La quota di idroelettrico pur essendo una fonte considerata continua (che quindi produce energia ininterrottamente a differenza di solare ed eolico) non è attualmente in crescita anche se sono previsti investimenti anche in questo ambito grazie al PNRR.

La restante parte del fabbisogno italiano è coperto da altre fonti rinnovabili come biocarburanti (biomasse, bioliquidi, biogas) e geotermico/geotermoelettrico[2].

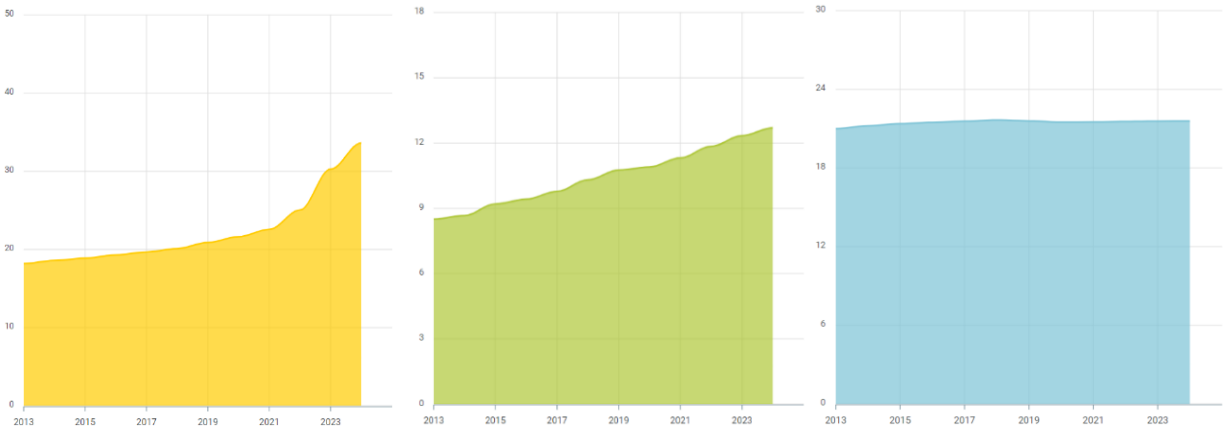

Figura 3: Produzione fotovoltaica (in giallo), eolica (in verde) ed idroelettrica (in blu) negli anni in Italia in GW, fonte: Terna

[1] Italia Domani: Piano Nazionale di Ripresa e Resilienza
[2] Terna: Dati da Dispacciamento-Fonti rinnovabili

3.4 Cosa è accaduto nel 2021

Nel 2021 l'Italia, e l'Europa in generale, ha vissuto un periodo di grandissima instabilità che ha portato anche ad uno sconvolgimento del mercato energetico italiano. In particolare, grazie alla ripresa economica post pandemia, a causa anche dei livelli attesi di estrazione di materie prime (come gas, carbone e silicio) non raggiunti, il prezzo di queste ha subito un forte rialzo. Il gas è stato quello che ha subito un aumento maggiore: passando da un valore medio di circa 15/20 €/MWh pre-ripresa economica, a toccare dei picchi di 180 €/MWh. Picchi, che con l'inasprimento dei rapporti tra UE e Russia a causa della guerra in Ucraina, hanno subito ulteriori rialzi toccando i 340 €/MWh[3].

Come visto dalla figura 2, più di un terzo della produzione energetica in Italia avviene tramite impianti termoelettrici a gas. Di questo gas, nel 2021, oltre il 45%[4] era importato dalla Russia; motivo per cui anche la produzione di energia elettrica ne ha risentito, facendo schizzare alle stelle anche il PUN: da valori medi di 50/60 €/MWh a picchi nel 2022 di 740 €/MWh[5]. Ad oggi, l'importazione di gas dalla Russia risulta quasi totalmente azzerata, scesa a meno del 2%, e sostituita con gas di altri paesi quali: Algeria, Azerbaijan, Qatar e GNL (gas naturale liquefatto) da più paesi[6].

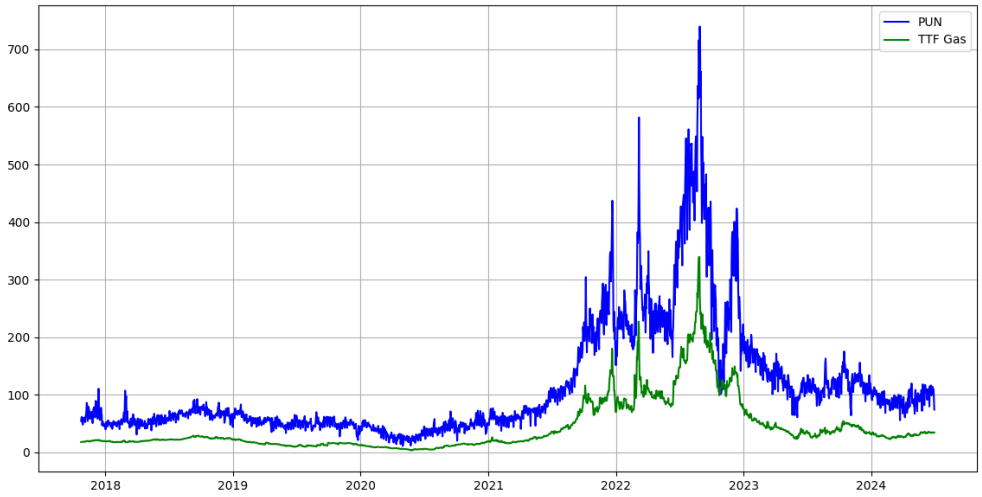

Figura 3: andamento del PUN e del prezzo del gas ('TTF Gas')

[3] Yahoo finanza: indice di riferimento europeo TTF gas (prezzo gas)
[4] Ministero dell'Ambiente e della Sicurezza Energetica (MASE): dati anno 2021
[5] GME: dati storici PUN
[6] SNAM: dati di esercizio

4 IL METODO

4.1 Modelli di previsione

Il costo dell'energia elettrica è un fattore chiave sia per i produttori che per i consumatori, e per sfruttare al massimo le opportunità offerte dal mercato dell'energia, i partecipanti devono essere in grado di prevedere i prezzi. Ecco perché l'uso di modelli predittivi avanzati può essere molto vantaggioso. Ci sono molti algoritmi in grado di fare previsioni; dalla semplice regressione, ad altri più avanzati algoritmi di machine learning che esploreremo in dettaglio. Nello specifico, in questa tesi cercheremo di costruire un modello di PUN basato sull'apprendimento automatico attraverso cross-validation.

4.1.1 Regressione – Lineare, Lasso e Ridge

La regressione lineare è uno dei metodi più semplici e intuitivi per prevedere i prezzi. Questo modello assume una relazione diretta tra le variabili indipendenti (fattori che influenzano il prezzo) e la variabile dipendente (il prezzo dell'energia). La formula generale è:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad [7]$$

dove Y rappresenta il prezzo dell'energia, $X_1, X_2, \dots, X_n$ sono le variabili indipendenti, β_0 è l'intercetta, $\beta_1, \beta_2, \dots, \beta_n$ sono i coefficienti delle variabili indipendenti, e ϵ rappresenta l'errore.

La regressione "*Lasso*" *(Least Absolute Shrinkage and Selection Operator)* è una variante che aggiunge una penalizzazione sui coefficienti, riducendo a zero quelli delle variabili meno significative. Questo aiuta a selezionare le caratteristiche più rilevanti:

$$Minimizza \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad [8]$$

dove λ è un parametro di penalizzazione che controlla quanto i coefficienti vengono ridotti.

[7] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*.
[8] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*.

La regressione Ridge, simile alla Lasso, introduce una penalizzazione che limita la magnitudine dei coefficienti, ma senza ridurli a zero. La formula di ottimizzazione per la regressione Ridge è:

$$Minimizza \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad [9]$$

In questo caso, la penalizzazione è basata sulla somma dei quadrati dei coefficienti, il che aiuta a prevenire l'*overfitting*, soprattutto quando ci sono molte variabili indipendenti collegate tra loro. Per *overfitting* si intende un problema comune in *machine learning* e statistica che si verifica quando un modello apprende "troppo bene" le particolarità del dataset di addestramento, inclusi i rumori e le variazioni casuali, a discapito della sua capacità di generalizzare su dati nuovi e non visti.

La regressione è un ottimo indicatore dell'andamento generale del modello, ma come suggerisce il nome, tende a seguire un andamento piuttosto "lineare", non riuscendo sempre a prevedere eventuali fluttuazioni improvvise.

4.1.2 Random Forest

Il modello *Random Forest* è un algoritmo di apprendimento automatico che utilizza un insieme di alberi decisionali. Combina molti alberi, ciascuno addestrato su un sottoinsieme casuale dei dati, per migliorare la precisione delle previsioni:

$$\hat{f}(X) = \frac{1}{B} \sum_{b=1}^{B} f_b(X) \qquad [10]$$

dove B è il numero di alberi nella foresta, f_b è il b-esimo albero, e $\hat{f}(X)$ è la previsione media dei singoli alberi.

La *Random Forest* è particolarmente efficace per gestire dati complessi e non lineari, riducendo il rischio di *overfitting* grazie al processo di aggregazione.

[9] Statology. "Introduction to Ridge Regression."
[10] Breiman, L. (2001). Random forests. *Machine Learning*

4.1.3 Gradient Boosting

Il *Gradient Boosting* è un metodo di *ensemble* (combina le previsioni di diversi modelli per migliorare la precisione e la robustezza rispetto ai singoli modelli) che costruisce modelli di previsione combinando deboli stimatori in una sequenza iterativa.

Ogni nuovo albero è addestrato per correggere gli errori commessi dagli alberi precedenti, migliorando progressivamente l'accuratezza del modello:

$$\hat{f}_m(X) = \hat{f}_{m-1}(X) + \eta \cdot g_m(X) \quad [11]$$

dove $\hat{f}_m(X)$ è la previsione del modello al passo m, η è il tasso di apprendimento, e $g_m(X)$ è l'albero addestrato sugli errori del modello precedente.

Questo algoritmo è potente per la sua capacità di adattarsi a relazioni non lineari e di gestire grandi volumi di dati. Tuttavia, richiede un'attenta sintonizzazione dei parametri per evitare l'*overfitting*.

4.1.4 XGBoost

Un'estensione del *Gradient Boosting* chiamata *XGBoost* mira a migliorare le prestazioni e la scalabilità dell'algoritmo. *XGBoost* migliora l'accuratezza complessiva combinando iterativamente le previsioni di modelli deboli (alberi di decisione) come *Gradient Boosting*. La formula di funzionamento è la medesima del *Gradient Boosting*, ma presenta l'aggiunta di un termine di regolarizzazione con l'obiettivo di prevenire l'*overfitting* controllando la complessità del modello. Inoltre, tecniche come il parallelismo nel calcolo e la gestione efficiente dei valori mancanti migliorano l'efficienza di *XGBoost*. Questo lo rende eccezionalmente adatto al lavoro con set di dati sparsi e grandi quantità di dati.

Tuttavia, per evitare l'*overfitting*, i parametri devono essere sintonizzati attentamente a causa della sua complessità. Anche se *XGBoost* è potente e può catturare relazioni non lineari complesse, la sua interpretabilità può essere inferiore rispetto ai modelli più semplici a causa delle numerose regolazioni e della struttura degli alberi.

[11] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*

4.1.5 Support Vector Regression (SVR)

Il *Support Vector Regression* (SVR) è un potente metodo di machine learning derivato dagli stessi principi del *Support Vector Machine* (SVM). SVR è progettato per eseguire la regressione, trovando una funzione che mappa *input a output* continui, mantenendo allo stesso tempo la complessità del modello sotto controllo.

SVR cerca di individuare una funzione $f(x)$ che ha al massimo una deviazione ϵ dalla vera risposta per tutti i punti del dataset, e allo stesso tempo è piatta quanto possibile. La formula di ottimizzazione è:

$$Minimizza\ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \max(0, |y_i - f(x_i)| - \epsilon) \quad [12]$$

Dove $\| w \|$ è la norma del vettore dei pesi, che SVR cerca di minimizzare per mantenere il modello semplice; C è un parametro di penalizzazione che determina il *trade-off* tra la dimensione della finestra di tolleranza ϵ e la complessità del modello. ϵ invece è una soglia che definisce una regione intorno alla funzione predetta entro la quale le predizioni sono considerate accettabili senza penalità.

SVR è particolarmente utile per problemi di regressione complessi dove la relazione tra variabili non è semplicemente lineare, essendo in grado di gestire dati non lineari e di prevenire l'*overfitting*.

[12] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004)

4.2 Valutazione dei dati

Per valutare le prestazioni dei modelli di previsione, i modelli saranno valutati prevedendo il prezzo in base a una serie di dati in entrata e confrontando poi l'output con il valore reale in quel momento. Tutti i dati ed i modelli saranno valutati con le stesse metriche.

4.2.1 Coefficiente di correlazione ($\rho_{X,Y}$)

Il coefficiente di correlazione viene utilizzato per determinare la correlazione relativa di due variabili dipendenti. La covarianza di due variabili viene divisa per il prodotto delle loro deviazioni standard per ottenere una correlazione percentuale tra le due variabili.

$$\rho_{X,Y} = \frac{\text{cov}\,(X,Y)}{\sigma_X \sigma_Y} \qquad [13]$$

$\rho_{X,Y}$ è sempre un valore compreso tra -1 e 1 per definizione. 1 equivale a una correlazione perfetta e -1 a una correlazione negativa perfetta. Per misurare la correlazione per la stessa serie temporale tra diversi punti nel tempo si possono utilizzare le funzioni di autocorrelazione. Tuttavia, l'obiettivo principale dell'analisi di correlazione in questa tesi è quello di comprendere la correlazione tra il prezzo dell'elettricità e le diverse variabili analizzate.

4.2.2 Indici di bontà di adattamento (R^2 e R^2 aggiustato)

L'R^2, o coefficiente di determinazione, rappresenta la proporzione della varianza nella variabile dipendente spiegata dalle variabili indipendenti del modello.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad [14]$$

Dove $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ è la somma dei quadrati degli errori di predizione (SSE), mentre $\sum_{i=1}^{n} (y_i - \bar{y})^2$ è la somma dei quadrati totali (SST).

L'R^2 aggiustato, invece, tiene conto del numero di predittori nel modello, offrendo una valutazione più precisa della bontà di adattamento, specialmente utile quando si confrontano modelli con diverso numero di variabili.

[13] James H. Stock, Mark W. Watson, "Introduzione all'econometria"
[14] James H. Stock, Mark W. Watson, "Introduzione all'econometria"

$$R^2_{\text{aggiustato}} = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right) \quad [15]$$

Dove n è il numero di osservazioni, p è il numero di predittori nel modello ed R^2 è il coefficiente di determinazione.

4.2.3 MSE e RMSE – Mean Squared Error e Root Mean Squared Error

Una misura molto comune in statistica e nell'analisi di regressione della qualità predittiva è l'errore quadratico medio (EQM o MSE). L'MSE calcola l'errore predittivo medio al quadrato di n campioni. Poiché l'errore è al quadrato, viene considerata la differenza relativa in ogni momento, vale a dire che una previsione eccessiva o insufficiente dà comunque un valore di MSE basso.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (A_t - F_t)^2 \quad [16]$$

Il RMSE è la radice quadrata del MSE e fornisce una misura della deviazione media delle predizioni dai valori osservati. Essendo nella stessa unità delle variabili di output, il RMSE è spesso più intuitivo da interpretare.

[15] James H. Stock, Mark W. Watson, "Introduzione all'econometria"
[16] P. Lara-Benitez´, M. Carranza-Garcia´, J. M. Luna-Romera, e J. Riquelme, "Temporal convolutional networks applied to energy-related time series forecasting," 2020

4.3 I dati

Per condurre un'analisi accurata e creare un modello solido è fondamentale un dataset di partenza completo e ben strutturato. I dati utilizzati in questo studio sono stati scaricati da diverse fonti ufficiali e autorevoli, assicurando l'affidabilità e la precisione delle informazioni. Di seguito una tabella riassuntiva dei dati utilizzati, tutti a frequenza giornaliera dal 24 ottobre 2017 (prima data disponibile per alcuni dei dati) al 30 giugno 2024:

Tabella 1: Dati utilizzati e fonti

Dati	*Fonte*	*Tipologia*	*Indicatore di prezzo*
PUN	GME	Dati storici	Prezzo medio
Materie prime	Investing.com	Futures	Prezzo di chiusura
Punto di Scambio Virtuale	Snam	Dati storici	Media giornaliera
Fabbisogno	Terna	Dati storici	Media giornaliera
Rischio geopolitico	Matteo Iacovello	Indice storico	Valore dell'indice
Emissioni di carbonio	Investing.com	Dati storici	Prezzo di chiusura
Dati climatici	Copernicus	Dati storici	-

4.3.1 PUN

I dati sul Prezzo Unico Nazionale sono stati scaricati dal sito di GME, autorità ufficiale italiana che si occupa della Gestione del Mercato Energetico e della raccolta di dati in tale ambito.

Dopo aver scaricato i file (uno per ogni anno a causa di limitazione del sito) in formato excel (.xlsx), è necessario prepararli per l'elaborazione. Grazie ad una serie di *script Python*, e tramite la libreria *Pandas* (libreria per la manipolazione e l'analisi dei dati) i dati sono stati convertiti in formato .csv, controllati per eventuali errori di formattazione e combinati in un unico file. Dopo aver unito i dati, sono stati puliti e preparati per l'analisi finale: rimozione di eventuali duplicati, correzione di errori nei dati e formattazione del *DataFrame* finale utilizzabile per le successive analisi.

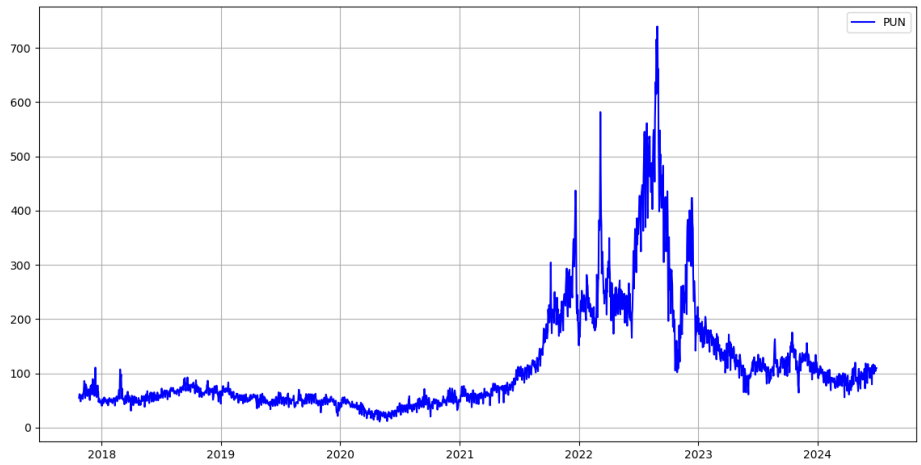

Figura 4: andamento del PUN in €/MWh

La figura 4 mostra l'andamento del PUN nel periodo di tempo considerato: risulta subito all'occhio la grande instabilità e l'impennata dei prezzi che si è verificata in seguito alla ripresa economica e allo scoppio della guerra in Ucraina.

4.3.2 Materie prime

I dati sulle materie prime sono stati scaricati tramite la piattaforma finanziaria *Investing.com.* Sempre con riferimento alla *figura 2* sono state selezionate le materie prime apparentemente più rilevanti: sicuramente gas e petrolio, considerando il fatto che insieme producono più della metà del fabbisogno energetico italiano. Anche il carbone è stato aggiunto all'analisi seppur la quota di energia prodotta da tale risorsa risulti esigua. Non avrebbe senso invece inserire nell'analisi anche l'uranio, visto che l'Italia attualmente non produce energia attraverso reattori nucleari.

Per quanto riguarda il gas, è stato selezionato il titolo "*Dutch TTF Natural Gas Futures*". Questo titolo rappresenta i contratti futures sul gas naturale scambiati sul *Dutch Title Transfer Facility (TTF)*, uno dei principali *hub* di *trading* di gas naturale in Europa. Per lo scopo dell'analisi è stato considerato solo il valore di chiusura. Dalla piattaforma Investing.com il primo dato disponibile risale al 23 ottobre 2017, motivo per cui tutti i dati usati avranno questa come data di inizio.

Per quanto riguarda il petrolio invece, esistendo due principali titoli, è stato selezionato il titolo "*Brent Oil"*. Questo viene utilizzato per quotare il greggio in Europa (inclusa la Russia), Africa e Medio Oriente, a differenza del titolo "*Crude Oil WTI"* che fa

riferimento al mercato in Nord e Sud America. Anche qui per lo scopo dell'analisi è stato considerato solo il valore di chiusura.

Per quanto riguarda il carbone, invece, sono due i titoli principali: "*Newcastle Coal Futures"* (rappresenta il prezzo del carbone scambiato presso il porto di Newcastle, in Australia) e "*Coal (API2) CIF ARA Futures"* (rappresenta il prezzo del carbone consegnato nei principali porti europei: Amsterdam-Rotterdam-Anversa). In questa analisi useremo "*Coal (API2)",* e considereremo anche qui solo il valore di chiusura.

Per quanto riguarda la preparazione dei dati, a differenza del PUN, è stato possibile scaricarli per tutto il periodo di interesse in un unico file excel. Utilizzando *Python e Pandas*, i dati sono stati convertiti in formato .csv, controllati e combinati in un unico file (avendo un file per ogni titolo). Essendo questi dei titoli, seguono l'andamento della borsa, che durante i fine settimana e le festività chiude. Pertanto, è stato necessario, tramite un'ulteriore elaborazione, riempire i valori mancanti del dataset: per farlo, è stato usato un codice Python con lo scopo di inserire, dove necessario, il dato del giorno precedente (così come accade per il valore dei titoli in borsa) riuscendo a colmare ogni valore mancante.

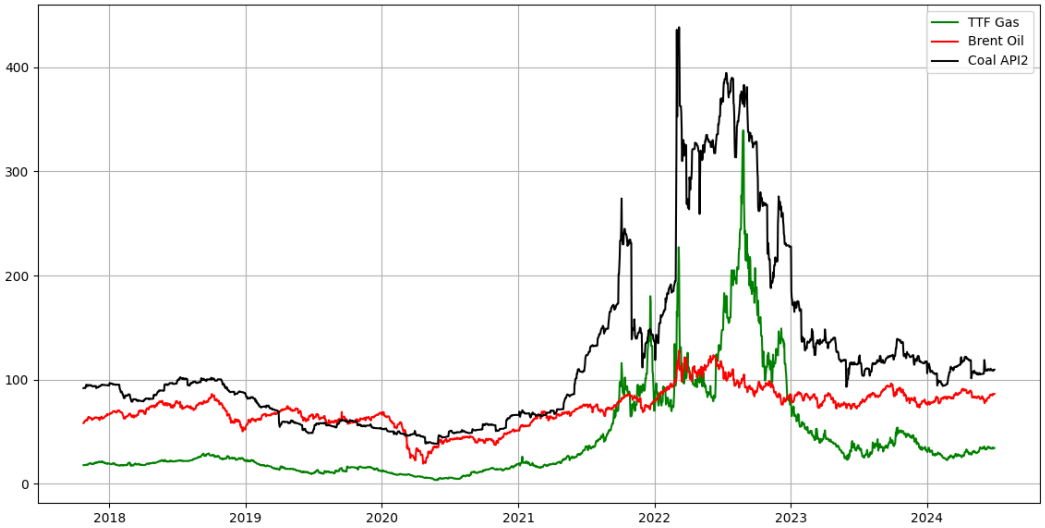

Figura 5: andamento di gas (TTF Gas), petrolio (Brent Oil) e carbone (Coal API2) negli anni in €/MWh

È interessante notare la somiglianza dell'andamento di gas e carbone rispetto al PUN (Fig. 4). Seppure i valori effettivi siano molto diversi, l'andamento è molto simile, anticipandoci una molto probabile buona correlazione. Il petrolio risulta la più stabile delle tre materie prime. A testimonianza del fatto che il greggio importato in Europa proviene solo in piccola parte dalla Russia, a differenza di gas e carbone, come già spiegato precedentemente.

4.3.3 Il Punto di Scambio Virtuale – PSV

Vista la grande correlazione del prezzo del gas con il PUN, può risultare utile aggiungere al *dataset* un ulteriore dato relativo al mercato del gas. L'indice PSV, che sta per "Punto di Scambio Virtuale", è un indice che riflette il prezzo all'ingrosso del gas naturale nel mercato italiano. Gestito da Snam Rete Gas, il PSV funziona come un hub virtuale dove produttori e fornitori si incontrano per compravendere il gas. Sostanzialmente rappresenta il prezzo del gas a livello nazionale, differenziandosi dall'indice TTF che si riferisce al mercato europeo in generale. Il prezzo al PSV, espresso in euro per metro cubo standard (€/Smc), viene aggiornato quotidianamente e riflette il prezzo "*Day Ahead*", ovvero il prezzo per il giorno successivo, che viene calcolato come media aritmetica dei prezzi giornalieri.

A differenza di quello che si possa pensare, però, gli indici PSV e TTF non sono uguali: i paesi di approvvigionamento non sono necessariamente i medesimi, portando a una differenziazione dei prezzi. Inoltre, l'indice TTF rappresenta il riferimento a livello europeo del prezzo del gas, causando indirettamente squilibri anche all'indice PSV.

Per quanto riguarda la preparazione dei dati, è stato possibile scaricare i dati dalla piattaforma di "segugio.it" riportante tutti i dati con cadenza giornaliera presi da Snam. Come per gli altri dati, anche questi sono stati convertiti in formato .csv, controllata la formattazione e riempiti i valori mancanti (tramite media dei valori del giorno precedente e quello successivo). I dati sono anche stati convertiti da €/smc a €/MWh sia per coerenza con le altre unità di misura riguardanti le materie prime, sia per ampliare la scala dei valori (essendo quella iniziale molto piccola).

La formula utilizzata è: $€/\text{MWh} = €/\text{Smc} \times \frac{\text{PCS}}{\text{fattore di conversione}}$

Dove "PCS" è il Potere Calorifico Superiore del gas impiegato, che è di circa 38 MJ/Smc per il metano (gas usato generalmente), mentre il fattore di conversione per passare da MJ a MWh è 3.6 (dato che 1 MWh = 3.6 MJ).

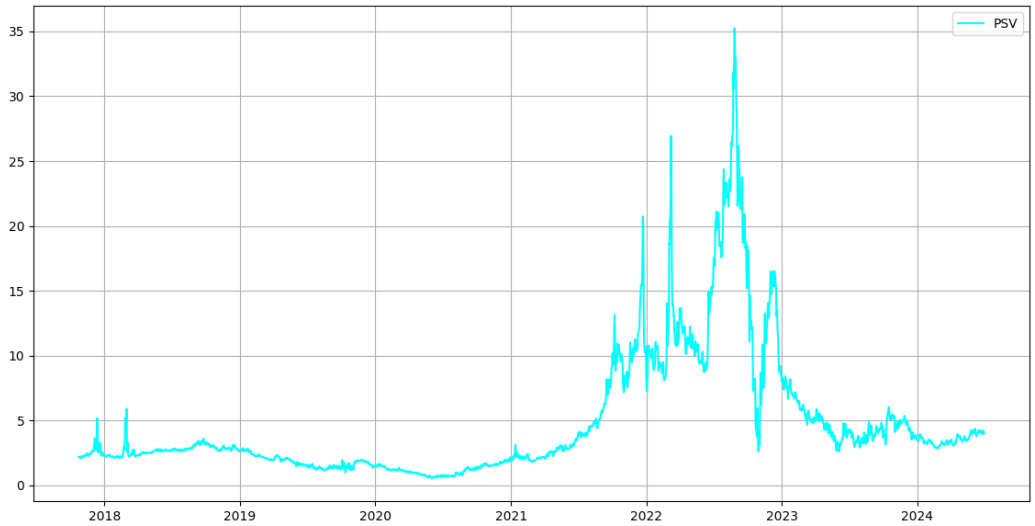

Figura 6: andamento dell'indice PSV

Nonostante la scala dei valori sia totalmente diversa, appare chiara una correlazione pressoché totale con il PUN, confermandoci la grande dipendenza dal gas per la produzione di energia elettrica.

4.3.4 Fabbisogno

I dati sul fabbisogno energetico sono stati scaricati dal sito di Terna.

Terna S.p.A. è una società italiana che gestisce la rete di trasmissione dell'energia elettrica ad alta tensione in Italia. È responsabile della pianificazione, sviluppo, gestione e manutenzione della rete elettrica nazionale e contribuisce alla sicurezza energetica del Paese. Raccoglie anche dati relativi a quasi ogni ambito relativo alla produzione e all'utilizzo dell'energia elettrica in Italia.

Dalla piattaforma "*Transparency report"* sono stati scaricati i dati della sezione *"load"* (fabbisogno appunto) dalla prima data disponibile fino al 30 giugno 2024. L'intento iniziale era quello di utilizzare dati sul rapporto domanda-offerta di energia elettrica in Italia. Tuttavia, non è stato possibile reperire dati scaricabili se non dal sito del GME, che permetteva il download dei dati un giorno alla volta.

Per quanto riguarda la preparazione dei dati, è stato possibile scaricarli per tutto il periodo di interesse in un unico file excel. Come per gli altri dati, anche questi sono stati convertiti in formato .csv e controllata la formattazione. I dati in questione erano scaricabili unicamente con intervalli di 15 minuti. Pertanto, è stato usato un ulteriore script Python

con lo scopo di sommare tutti i valori riferiti ad ogni giorno nel dataset, così da ottenere i valori del fabbisogno giornalieri.

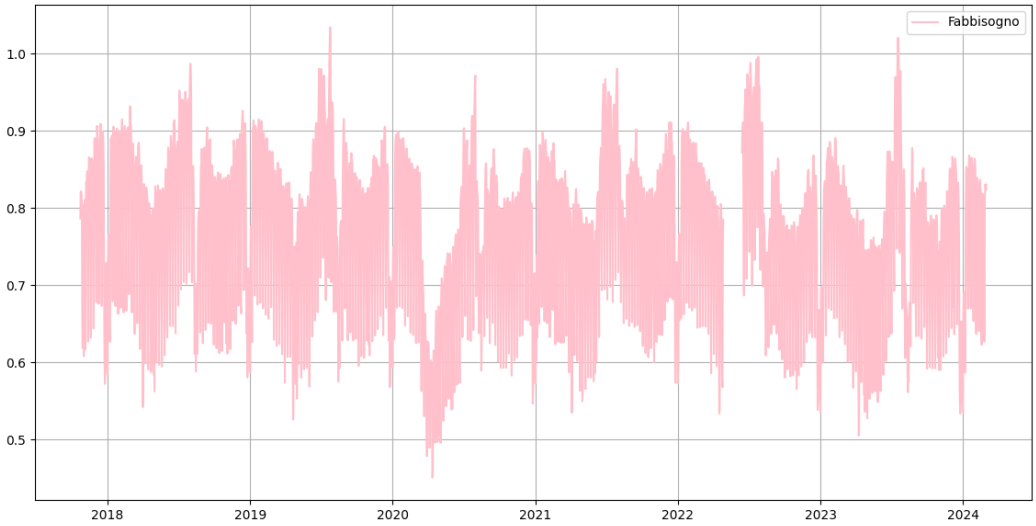

Figura 7: Andamento del fabbisogno in milioni di GW

Da una veloce analisi del grafico si nota come il fabbisogno segua la stagionalità, con valori più alti in estate (per l'aria condizionata) e in inverno (per il riscaldamento). Si nota anche l'effetto del lockdown causato dal Covid, con un picco negativo nel secondo quadrimestre del 2020 che ha portato alla chiusura di quasi tutte le industrie e le attività commerciali. Si evince inoltre la poca somiglianza con l'andamento del PUN, indicandoci una probabile bassa correlazione.

4.3.5 Indice di rischio geopolitico (GPR – GeoPolitical Risk Index)

Vista la grande instabilità degli ultimi anni e di come soprattutto il prezzo dell'energia elettrica sia variato così tanto a causa di diversi eventi geopolitici, può risultare interessante aggiungere un dato di questo tipo. Essendo la maggior parte degli indici GPR non di pubblico dominio (es. BlackRock) è stato usato un indice disponibile per tutti, creato da Dario Caldara e Matteo Iacoviello.

L'indice in questione è una misura degli eventi geopolitici avversi e dei rischi associati sulla base di un conteggio degli articoli di giornale che coprono le tensioni geopolitiche, e ne esaminano l'evoluzione e gli effetti economici dal 1900. L'indice registra picchi in corrispondenza delle due guerre mondiali, dell'inizio della guerra di Corea, durante la crisi dei missili di Cuba e dopo l'11 settembre. L'indice utilizza le notizie di 3 testate

giornalistiche per l'indice storico dal 1900 fino al 1985, mentre ne utilizza 10 per quello moderno.

Un rischio geopolitico più elevato fa presagire una riduzione degli investimenti, dei prezzi delle azioni e dell'occupazione. È anche associato a una maggiore probabilità di disastri economici e a maggiori rischi di ribasso per l'economia globale.

Per quanto riguarda la preparazione dei dati, è stato possibile scaricare il file excel con tutti i dati: sia quelli sull'indice storico, sia quelli riferiti ai singoli paesi con cadenza giornaliera. Come per gli altri dati, anche questi sono stati convertiti in formato .csv, controllata la formattazione e selezionata solo la colonna utilizzata (quella dell'indice "generico").

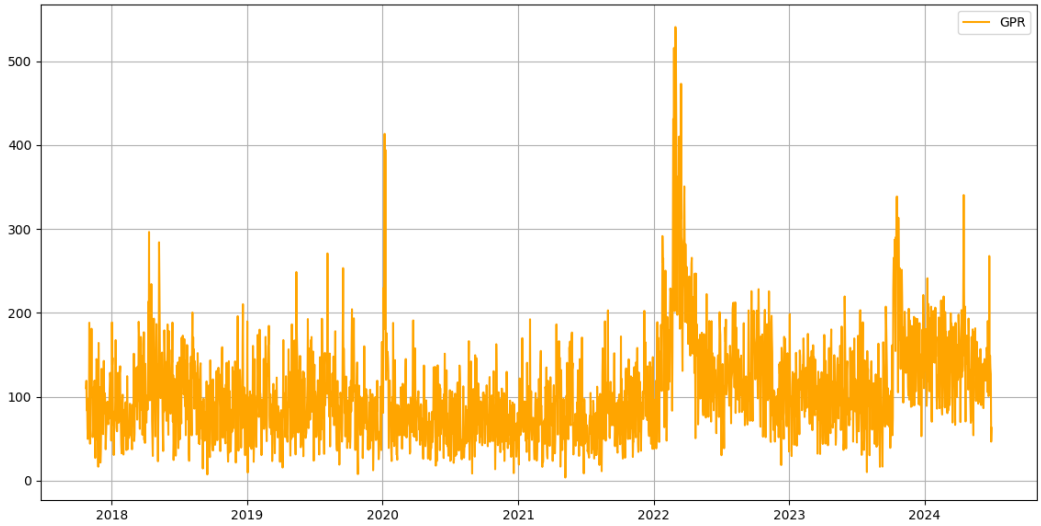

Figura 8: Andamento dell'indice GPR

Il grafico mostra un andamento simile a quello del PUN nel periodo 2021-2022, con un aumento significativo e un picco nel 2022, seguito da una riduzione e maggiore variabilità. La correlazione tra i due grafici può indicare che le circostanze del mercato influenzano sia il GPR che il PUN contemporaneamente. Tuttavia, è difficile determinare una correlazione dalla sola analisi visiva.

4.3.6 EU ETS ed Indice di Emissioni di Carbonio

Vista la grande quota di energia prodotta in Italia da fonti non rinnovabili, potrebbe essere interessante verificare l'impatto delle regolamentazioni europee in ambito di emissioni di CO_2.

Il sistema EU ETS (*European Union Emission Trading System*) si basa sul principio del "cap and trade", dove viene fissato un limite massimo (cap) sulle emissioni di gas serra degli impianti e dagli operatori aerei che rientrano nel sistema. Questo limite è ridotto annualmente per garantire una diminuzione delle emissioni nel tempo. Le aziende devono acquistare permessi (quote) per emettere CO_2, ricevendone anche alcune gratuitamente, e possono scambiarle tra loro. Se riducono le emissioni, possono vendere o conservare le quote in eccesso. Il sistema genera ricavi dalla vendita delle quote, che gli Stati membri utilizzano per finanziare investimenti in energie rinnovabili e tecnologie a basse emissioni di carbonio.

Per il dataset useremo il Carbon Emissions Index, che è lo strumento finanziario che riflette il prezzo dei diritti di emissione di CO_2.

Per quanto riguarda la preparazione dei dati, il procedimento è stato analogo a quello delle materie prime, avendo scaricato anche questo dato dalla piattaforma Investing.com: utilizzando *Python e Pandas*, il file excel è stato convertiti in formato .csv, controllata la formattazione e riempiti i valori mancanti (inserendo dove necessario il valore del giorno precedente come per gli altri titoli finanziari).

Figura 9: andamento dell'indice EU ETS (sulle emissioni di carbonio) in €/tonnellata di CO_2 emessa

Nonostante la scala dei valori sia notevolmente diversa rispetto al PUN (EU ETS oscilla tra 10 e 100, mentre il PUN varia tra 40 e 700), anche in questo grafico risulta chiaro il picco del 2021-2022, suggerendo un possibile collegamento con il PUN, anche se l'intensità della discesa e la stabilizzazione successiva differiscono.

4.3.7 Dati climatici

In un paese con una buona produzione di energia da fonti rinnovabili, ma anche solo in un paese con temperature più o meno rigide può essere utile aggiungere all'analisi qualche dato relativo al clima per capire se, ad esempio, la produzione o il consumo di energia possa essere influenzata dalle temperature, dal vento o dalle precipitazioni.

Non avendo trovato alcun dataset pronto con questi dati, sono stati scaricati i dati grezzi direttamente dall'archivio *ERA5 di Copernicus*. ERA5 è la quinta generazione *dell'European Centre for Medium-Range Weather Forecasts* (ECMWF) per la rianalisi atmosferica, fornisce stime orarie per un gran numero di variabili climatiche atmosferiche, terrestri e oceaniche.

Le variabili scelte sono:

1) t2m: temperatura a 2 metri dal suolo.
2) ssr / ssrd / ssrdc: irraggiamento solare al suolo con cielo nuvoloso e non.
3) u10 / u100 / v10 / v100: velocità del vento a 10 e 100 metri di quota; dove u e v indicano la direzione (u= est/ovest, v= nord/sud).
4) tp: precipitazioni totali.

Dopo aver fatto richiesta dei dati, dopo circa 24 ore erano disponibili. I dati erano in formato NetCDF, formato complesso che si presta bene per immagazzinare dati multidimensionali, a differenza del più semplice e facile da usare formato .csv capace di supportare dati al massimo bidimensionali. I file in questione erano tridimensionali (latitudine, longitudine e tempo), infatti contenevano i dati sopra menzionati con cadenza oraria e con intervalli spaziali di 0.5 gradi sia verticali che orizzontali per tutta l'area selezionata.

La selezione dell'area era possibile solo tramite selezione di una "regione" del mondo fornendo i limiti di latitudine e longitudine, nel nostro caso: Nord: 47.08; Ovest: 6.67; Sud: 36.67; Est: 18.66 che corrispondono circa all'area mostrata in figura 10.

Figura 10: Area selezionata per i dati metereologici, fonte:Google Maps

I file (uno per ogni anno per limiti della piattaforma) sono stati convertiti in più .csv (uno per ogni variabile) grazie ad uno script python[17] della guida ufficiale ECMWF. Dopodiché uniti in un unico dataset e mediati per ottenere alla fine dati medi giornalieri per l'intervallo di tempo considerato.

Anche qui si nota, ovviamente, come i dati seguano la stagionalità, con valori più alti in estate e più bassi in inverno (per quanto riguarda l'irraggiamento e la temperatura) e viceversa per il vento.

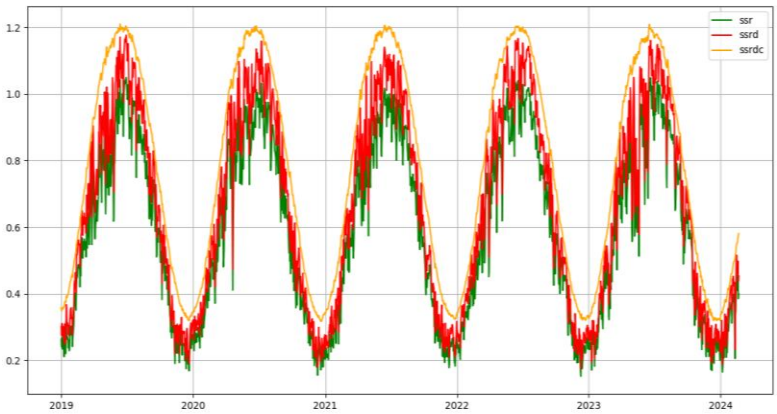

Figura 12:Irraggiamento solare in milioni di "J/m²"(Joule su metro quadro)

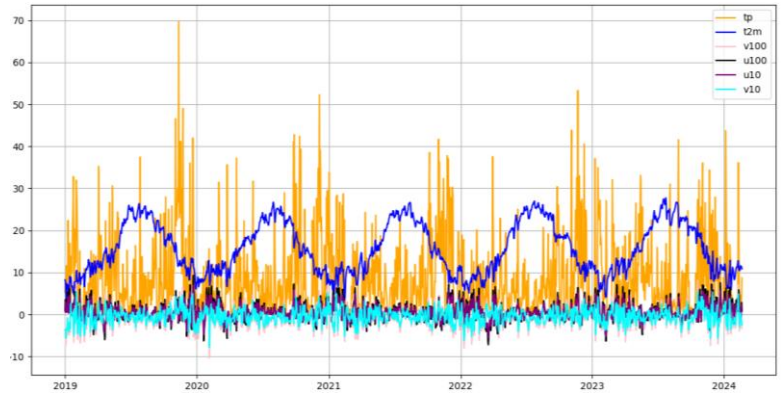

Figura 12: velocità del vento in "m/s", temperatura in "°C", e precipitazioni in "cm"

[17] https://confluence.ecmwf.int/display/CKB/How+to+convert+NetCDF+to+CSV

4.3.7.1 Considerazioni

È doveroso far notare che i dati utilizzati, nonostante provengano da una fonte autorevole quale l'ESA (agenzia spaziale europea), non siano comunque ottimali per questa analisi (a meno di altri importanti elaborazioni di dati), visto che è stato possibile scaricare dati solo tramite selezione di un'area dalla mappa. Difatti, i dati non sono altro che la media dei valori corrispondenti a tutte le possibili coppie di coordinate all'interno dell'area selezionata: ma in questa figurano anche zone al difuori della nostra area di interesse. Infatti, come si evince dalla figura 10, sono incluse anche aree della Tunisia, dell'Austria, della Bosnia e dell'Ungheria; oltre che quasi la totalità della Svizzera, della Croazia, della Slovenia e un'area piuttosto estesa di Mar Mediterraneo. Queste aree hanno un clima abbastanza diverso da quello italiano, falsando in parte i dati.

Generalmente, però, l'irraggiamento solare e la velocità del vento influenzerebbe il prezzo dell'energia attraverso la loro rispettiva produzione di energia rinnovabile. In questo caso però, le variabili meteorologiche analizzate, apparentemente, mostrano una correlazione molto debole con il PUN. Questo potrebbe essere spiegato, probabilmente, dalla diversità del clima e dalle fonti di generazione di energia elettrica in Italia. Il clima differisce molto tra nord e sud Italia: sono piuttosto comuni le giornate molto soleggiate e afose al sud con temporali al nord ad esempio. Questo causerà sicuramente un grande squilibrio nella produzione solare tra sud e nord.

Inoltre, come già ampiamente spiegato, l'Italia basa ancora la maggior parte della sua produzione di energia elettrica da fonti non rinnovabili, motivo per cui anche se gli impianti solari e eolici producessero energia a pieno regime per un lungo periodo, non provocheranno mai un impatto significativo al PUN.

4.4 *Cross validation*

L'idea di questa tesi è quella di creare un modello di previsione e verificarne l'accuratezza con la *cross validation*. È una tecnica che consiste nel dividere il dataset in più parti (chiamate "*fold*"), allenare il modello su alcune di queste parti e testarlo sulle restanti, ripetendo il processo più volte. Questo permette di ottenere una stima più affidabile delle prestazioni del modello, riducendo il rischio di *overfitting* e valutando la sua capacità di generalizzare su dati non visti. Nel nostro caso però non avrebbe senso usarla nella sua forma "standard", visto che, essendo il dataset una serie temporale, andrebbe ad allenare il modello sia con dati passati che futuri, rendendo il modello "inutile" e ottenendo risultati inaccurati.

Figura 13: cross validation "standard", fonte: mathworks

Motivo per cui è stato utilizzato un approccio di "*rolling window*". Questo metodo è una variante della *cross-validation* che consente di allenare e convalidare il modello su diversi periodi di tempo, testandone poi le prestazioni su un periodo specifico non visto in precedenza. Questo permette una valutazione più robusta del modello su differenti periodi temporali.

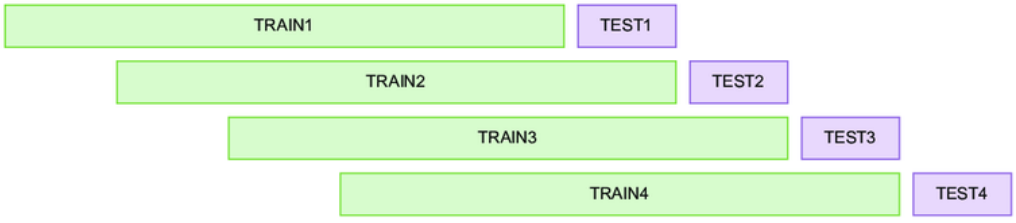

Figura 14: cross validation - rolling window, fonte: Reseach Gate

5 RISULTATI

5.1 Statistiche descrittive

Le statistiche descrittive forniscono una panoramica sulle caratteristiche principali dei dati.

Tramite un semplice script *python* (tramite la funzione *describe* della libreria *Pandas*) si ottengono le statistiche descrittive di tutte le variabili, che per comodità verranno separate in due tabelle. I risultati ottenuti includono: media, dati sulla dispersione (deviazione standard e varianza), il coefficiente di variabilità (ottenuto da: $\frac{\text{Deviazione Standard}}{\text{Media}}$), minimo e massimo per ogni variabile.

Tabella 2: Statistiche descrittive 1, dove STD= deviazione standard e CV= coeff. di variabilità

	PUN	TTF GAS	BRENT OIL	GPR	COAL API2	PSV	EU ETS	LOAD
MEDIA	113,6461	42,35111	70,93552	103,1584	117,8126	4,718375	46,49672	775010,9
STD	105,857	47,54603	17,5651	58,30345	81,33649	5,06535	27,91767	103638,4
CV	**0,93**	**1,12**	**0,25**	**0,56**	**0,69**	**1,07**	**0,60**	**0,13**
MIN	10,80514	3,51	19,33	3,565735	38,6	0,523913	7,55	451170,7
MAX	739,6456	339,2	127,98	540,8274	438,35	35,23974	98,01	1034137

Tabella 3: Statistiche descrittive 2

	SSR	SSRD	SSRDC	U10	U100	V10	V100	T2M	TP
MEDIA	589615,24	663378,27	779325,53	0,63	0,81	-0,43	-0,47	15,67	8,91
STD	265694,68	294742,13	311138,27	1,67	2,18	1,73	2,33	6,03	8,90
CV	**0,45**	**0,44**	**0,40**	2,67	2,70	-4,02	-4,91	**0,38**	**1,00**
MIN	152310,99	174475,17	316913,20	-5,53	-7,21	-7,78	-10,36	3,42	0,01
MAX	1049288,41	1176515,58	1209758,56	6,97	9,35	5,50	8,09	27,75	69,70

I risultati di questa analisi rivelano diverse informazioni chiave sulle variabili del dataset, soprattutto per quanto riguarda la volatilità dei valori in generale:

Il PUN, così come entrambe le variabili sul Gas, mostrano una notevole variabilità, con un coefficiente di variabilità circa uguale a 1.

Questo implica prima di tutto un'elevata variabilità: i dati sono altamente variabili rispetto alla media. Ad esempio, se consideriamo la variabile "PUN", la media è circa 113 mentre

la deviazione standard è 105: questo significa che i prezzi possono variare ampiamente intorno alla media, con alcune osservazioni significativamente più alte o più basse di 105 rispetto al valore della media (se la media è di 113, l'intervallo di valori è: 8/218).

Questo è anche confermato dai valori minimi e massimi registrati: la variabile PUN, ad esempio, contiene valori da 10 fino ad oltre 700. Lo stesso discorso è valido per quasi tutte le altre variabili (anche se alcune su scale di valori differenti).

In contesti finanziari o economici, un CV di 1 potrebbe indicare un alto livello di rischio e incertezza. Ad esempio, un investimento con un rendimento medio pari alla sua deviazione standard sarebbe considerato molto rischioso.

Questa variabilità risulta decisamente meno impattante per le altre variabili, in particolare per il fabbisogno ed i dati climatici: per quanto riguarda i dati sul vento (che riportano valori tra 2 e 4), il coefficiente non risulta veritiero. Questo per la natura stessa dei dati: la variabile è formata sia da dati positivi che negativi, motivo per cui il valore del CV può risultare fuorviante.

Fatte queste considerazioni risulta innegabile la complessità dei dati, e di conseguenza la difficoltà nella creazione di un modello predittivo molto accurato. Per quanto riguarda R^2 non è possibile stabilire quale sia un valore "accettabile". Per quanto riguarda l'RMSE, invece, è possibile basarsi sulla deviazione standard: un RMSE che rappresenta una piccola frazione della deviazione standard del PUN (ad esempio, meno del 30%) può essere considerato buono. Nel caso del PUN, la deviazione standard è 105.85, quindi un RMSE inferiore a circa 30 può essere considerato soddisfacente.

5.2 Correlogramma

Questo paragrafo illustra i risultati dell'analisi di correlazione delle variabili analizzate. Per facilitarne la visualizzazione i risultati sui dati meteo saranno inseriti in un secondo correlogramma. Le correlazioni indicate sono calcolate tramite coefficiente di correlazione.

I prezzi dell'elettricità sono denominati con "PUN". I contratti future per i prezzi delle materie prime sono denominati con il rispettivo nome: "Gas", "Coal" o "Oil". L'indice di rischio geopolitico, così come il Punto di Scambio Virtuale, sono abbreviati

rispettivamente con “GPR” e “PSV”. Il fabbisogno con “Load”, mentre l’indice di emissione di carbonio con il più generico “EU ETS”.

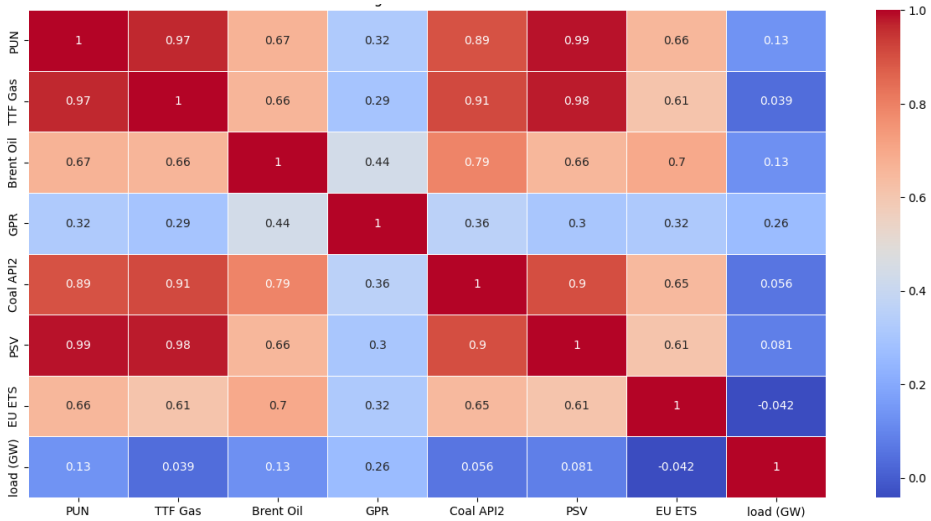

Figura 15: correlogramma delle variabili

Per quanto riguarda i dati climatici, sono stati comparati solo con il PUN per facilitarne l’interpretazione. Ogni variabile è stata chiamata come sopra indicato (nel paragrafo 4.3.7).

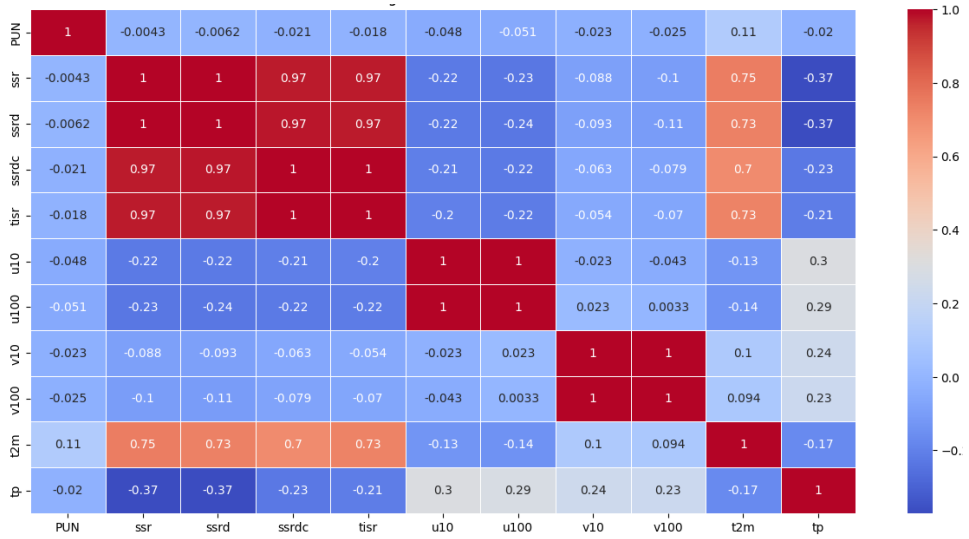

Figura 16: Correlogramma delle variabili climatiche

Con questi risultati si sono verificate le ipotesi precedenti. Le materie prime hanno un impatto significativo sul PUN, in particolare gas e carbone. Anche l’indice di emissione di CO_2 ha una buona correlazione nonostante l’andamento post ripresa economica molto diverso. L’indice GPR ha una modesta correlazione, motivo per cui bisognerà valutare (così come per “EU ETS”) se mantenerlo nel dataset o meno. Inaspettatamente i risultati del fabbisogno e dei dati climatici non sono buoni, avendo ottenuto una correlazione molto bassa; motivo per cui sono stati esclusi dal dataset senza ulteriori analisi.

5.3 Scatterplot

Di seguito viene riportato lo scatterplot delle variabili, avendo già escluso i dati climatici. Uno scatterplot, o diagramma di dispersione, è un tipo di grafico utilizzato per visualizzare la relazione tra due variabili quantitative. Non è altro che una rappresentazione cartesiana di tutte le possibili coppie di dati del dataset. Risulta utile per identificare eventuali andamenti lineari tra le variabili.

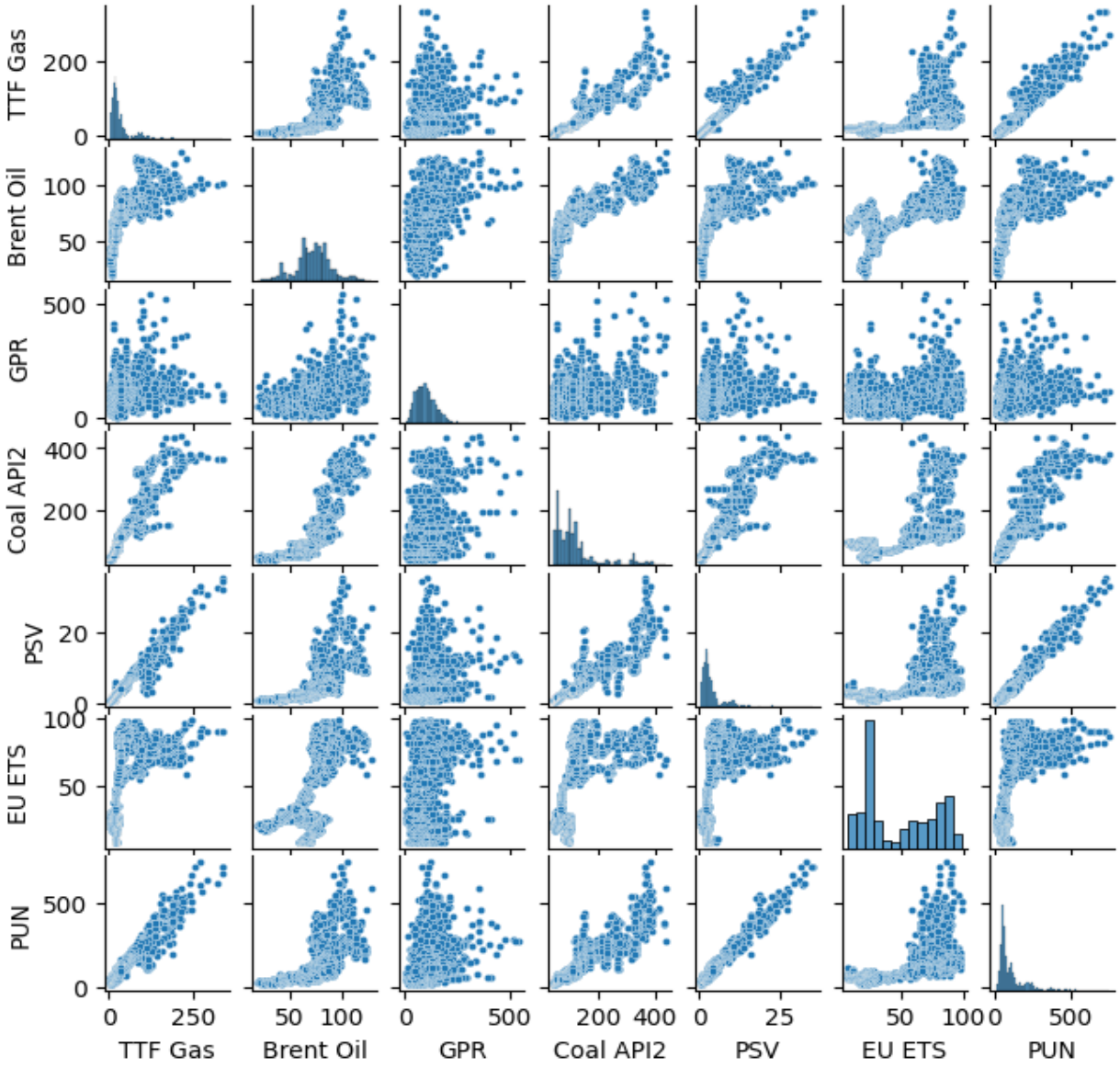

Figura 17: Scatterplot delle variabili

Dallo scatterplot innanzitutto è possibile confermare le conclusioni tratte con il correlogramma: il gas ed il carbone sono altamente correlate con il PUN mostrando quasi una retta. Per quanto riguarda invece gli indici di rischio geopolitico e di emissioni di carbonio, si notano nuvole di punti decisamente più "disordinate" rispetto alle altre

variabili. Motivo per cui testeremo il modello sia con che senza queste variabili per trovare la combinazione ideale.

Risulta anche interessante notare (come già precedentemente anticipato) la correlazione molto alta tra alcune delle variabili indipendenti, ad esempio: “PSV” e “TTF Gas”, “Coal API2” e “TTF Gas”. Questo potrebbe causare multicollinearità nel modello: per multicollinearità si intende quel fenomeno statistico che si verifica quando due o più variabili indipendenti sono altamente correlate tra loro. Questo significa che una variabile può essere quasi perfettamente predetta dalle altre variabili indipendenti presenti nel modello.

Questo porterebbe innanzitutto ad un’instabilità delle stime dei coefficienti (che possono diventare molto sensibili ai piccoli cambiamenti nei dati), oltre a delle varianze elevate (rendendo meno affidabili i test di significatività statistica).

Per ovviare al problema è stata inserita nel modello la “ridge regression” che, come già accennato, utilizza metodi di regressione penalizzata, che possono gestire la multicollinearità aggiungendo un termine di penalizzazione ai coefficienti.

5.4 Alcune considerazioni sui dati

Analizzando questi primi risultati emergono alcuni fatti che meritano di essere approfonditi. Una delle principali osservazioni emerse riguarda la differenza di impatto tra gas e petrolio sul PUN, nonostante entrambe le fonti energetiche abbiano una quota simile nella produzione di energia (figura 2). A seguito di alcune ricerche, due sono le ipotesi più plausibili rinvenute: la prima riguarda il fatto che le centrali elettriche alimentate a gas operino frequentemente come centrali di carico di base, funzionando in modo continuativo per assicurare una fornitura stabile di energia. Questo fa sì che le variazioni nel prezzo del gas si riflettano quasi immediatamente sul PUN. Le centrali a petrolio, al contrario, sono generalmente utilizzate solo in momenti di picco della domanda o in situazioni di emergenza, il che fa sì che il prezzo del petrolio influenzi il PUN in maniera meno frequente e diretta rispetto al gas.

La seconda ipotesi riguarda il fatto che il prezzo del gas è particolarmente sensibile a eventi geopolitici, come dimostrato dalla crisi tra Russia e Ucraina, che ha provocato forti oscillazioni nei prezzi a causa della dipendenza dell'Europa dalle forniture di gas russo.

A differenza del petrolio, che essendo più facilmente trasportabile, le fonti di approvvigionamento possono essere molto più differenziate, rendendolo molto meno suscettibile a variazioni dovute a crisi geopolitiche.

Un'altra considerazione riguarda il carbone che, pur rappresentando una quota molto piccola nella produzione energetica complessiva, esercita un impatto significativo sul PUN. Questo è probabilmente dovuto ai costi elevati associati alle regolamentazioni sulle emissioni di CO_2, che impongono alle centrali a carbone di acquistare permessi di emissione, incrementando così i costi di produzione e incidendo sul prezzo finale dell'energia. Come dimostrato anche dalla correlazione molto alta tra carbone e indice di emissione di CO_2 (EU ETS).

5.5 Dati utilizzati

Dopo diversi test eseguiti, sono state escluse dal modello quelle variabili che non contribuivano in modo significativo al miglioramento delle previsioni del modello: l'indice di emissione di carbonio (EU ETS) ed il Fabbisogno (Load).

L'esclusione di queste variabili è stata basata su una serie di test iterativi, dove ogni variabile è stata aggiunta e rimossa per osservare il cambiamento nelle performance del modello. Questo processo ha permesso di identificare le variabili più influenti e di costruire un modello più efficiente e accurato per la previsione del PUN.

5.6 Cross validation

In questa fase del progetto si proverà prima di tutto ad addestrare un modello predittivo utilizzando diversi algoritmi, per poi verificarne le prestazioni tramite cross validation (rolling window) e, per semplicità e facilità di interpretazione, due soli indicatori di performance: RMSE ed R^2.

5.6.1 Regressione Ridge

Tutto il processo di creazione del modello è avvenuto su *python* sfruttando la libreria *Scikit-learn*. Prima di tutto, una parte del codice si occupa di caricare e preparare i dati.

Si passa poi alla fase di *preprocessing* dei dati: la colonna della data viene rimossa ed i valori vengono standardizzati. La standardizzazione serve per uniformare la scala dei valori delle variabili del dataset, migliorando le prestazioni degli algoritmi. In seguito, si definiscono le caratteristiche (*features*) e il target (*label*) del modello. Nel nostro caso, le caratteristiche sono tutte le colonne eccetto 'PUN', che è la variabile di interesse.

Dopodiché deve essere impostato il modello da usare. In questo caso, il primo algoritmo testato è la regressione Ridge, che come spiegato in precedenza è una variante della regressione lineare che include un termine di penalizzazione per ridurre il rischio di *overfitting*.

Infine, il codice esegue la *cross-validation* (con approccio *rolling window*) e stampa i punteggi ottenuti per ciascuna delle "sezioni" definite (nel nostro caso 5). I punteggi calcolati non sono altro che gli R^2 e di RMSE di ogni *split*. Viene calcolata anche la media degli indicatori, fornendo una misura complessiva delle prestazioni del modello.

L' R^2 medio ottenuto è di circa 0.71. In un modello di regressione significa che il modello spiega il 71% della variabilità dei dati di risposta attorno alla media, risultando in un modello discreto ma migliorabile. Questo è un risultato molto buono che ci si può aspettare in questa fase visto che, come già spiegato, la regressione lineare non si presta bene in caso di grande variabilità dei dati (come nel caso del PUN). RMSE ottenuto invece è di circa 23, risultato abbastanza soddisfacente, il quale indica che in media i risultati ottenuti deviano dai valori effettivi di 23 (€/MWh).

5.6.2 Random Forest

Analogamente a quanto fatto per la regressione *Ridge*, allo stesso modo è stato testato il modello di *Random Forest*, che dovrebbe comportarsi meglio in caso di andamento non lineare. L' R^2 medio ottenuto è di 0.25 circa, mentre l'RMSE è di 55 circa, indicando un deciso peggioramento nella precisione della previsione.

Per ottimizzare ulteriormente le prestazioni del modello, è stata implementata anche un'ottimizzazione dei modelli, tramite "*GridSearch*", che esplora varie combinazioni di *iperparametri* per trovare la configurazione migliore (nei paragrafi successivi verrà spiegata più approfonditamente). In questo caso il migliore R^2 ottenuto è stato di 0,31, mentre l'MRSE di 52 incrementando leggermente il risultato precedente.

5.6.3 Gradient Boosting, XGBoost e SVR

Il modello è stato testato anche con il *Gradient Boosting Regressor, XGBoost e SVR*. Allo stesso modo della *Random Forest*, per ottimizzare le prestazioni è stata utilizzata la *GridSearch* per trovare la configurazione migliore. I risultati ottenuti sono interessanti: per quanto riguarda *Gradient Boosting e XGBoost*, gli R^2 ottenuti sono stati rispettivamente di 0.38 e 0.19, mentre gli RMSE di 45 e 61. Per quanto riguarda *SVR*, i risultati sono stati decisamente migliori, ottenendo un R^2 medio di 0.68 e un RMSE di 24.

Tabella 4: confronto algoritmi prima e dopo l'ottimizzazione degli iperparametri

	ALGORITMI NON OTTIMIZZATI		ALGORITMI OTTIMIZZATI	
	R^2	***RMSE***	***R^2***	***RMSE***
RIDGE	0.71	23	-	-
RANDOM FOREST	0.25	55	0.31	52
GRADIENT BOOSTING	0.40	49	0.38	45
XGBOOST	0.19	61	0.19	61
SVR	-1.01	103	0.68	24

Interessante notare come a seguito dell'ottimizzazione ci siano stati miglioramenti, (anche sostanziali) per quanto riguarda sia Random Forest che SVR, mentre per altri algoritmi non c'è stato nessun cambiamento significativo: nel caso di *XGBoost*, ad esempio, i migliori parametri risultano quelli "standard" (motivo per cui verranno mantenuti i parametri base), mentre nel caso di *Gradient Boosting*, si è ottenuto sia un R^2 minore, sia un RMSE minore. Il modello Ridge non è stato ottimizzato in quanto non fondamentale farlo, anche visti gli ottimi risultati ottenuti dai parametri standard. Nel paragrafo seguente verranno mostrati tutti i parametri utilizzati.

5.6.4 Gli iperparametri

La selezione degli iperparametri per modelli di apprendimento automatico di grandi dimensioni è molto complessa a causa del numero di combinazioni possibili. La selezione è stata quindi semplificata e affidata a *GridSearch*, i cui output si basano esclusivamente sui risultati migliori di R^2 e RMSE, oltre che sul tempo di addestramento.

Di seguito una tabella riassuntiva degli iperparametri migliori trovati con la spiegazione dei parametri.

Tabella 5: Migliori Iperparametri identificati

Modello	***Parametro***	***Valore***	***Spiegazione***
Regressione Ridge	-	-	Nessuna ottimizzazione è stata effettuata per questo algoritmo
Random Forest	min_samples_leaf	1	Numero minimo di campioni richiesti per essere un nodo foglia.
	min_samples_split	5	Numero minimo di campioni richiesti per dividere un nodo.
	n_estimators	100	Numero di alberi nella foresta.
	random_state	42	Seed utilizzato dal generatore di numeri casuali per garantire la riproducibilità dei risultati.
	max_depth	7	Profondità massima degli alberi nel modello.
Gradient Boosting	learning_rate	0.1	Tasso di apprendimento che determina l'impatto di ciascun albero aggiunto.
	max_depth	7	Profondità massima degli alberi nel modello.
	n_estimators	50	Numero di alberi nel modello di boosting.
	subsample	0.8	Frazione di campioni utilizzata per addestrare ciascun albero.
	random_state	42	Seed per il generatore di numeri casuali.
XGBoost	-	-	L'ottimizzazione non ha portato miglioramenti rispetto al modello senza ottimizzazione
SVR	C	10.0	Parametro di regolarizzazione che controlla il trade-off tra errore di classificazione e margine.
	degree	3	Grado del polinomio utilizzato nel kernel se `kernel='poly'.
	epsilon	0.01	Margine in cui nessun errore è penalizzato nella funzione di perdita epsilon-insensitive.
	kernel	linear	Funzione kernel utilizzata per trasformare i dati.

5.7 *Confronto dei risultati*

Per quanto riguarda la selezione degli algoritmi da inserire nel dataset, è importante analizzare accuratamente i risultati prima ottenuti: per fare questo e per rendere la spiegazione il più chiaro possibile, verrà sfruttato un grafico di confronto con dei *box plot.*

In questo grafico vengono confrontate le prestazioni di ciascun algoritmo in termini di R^2, comprensivi dell'ottimizzazione degli iperparametri.

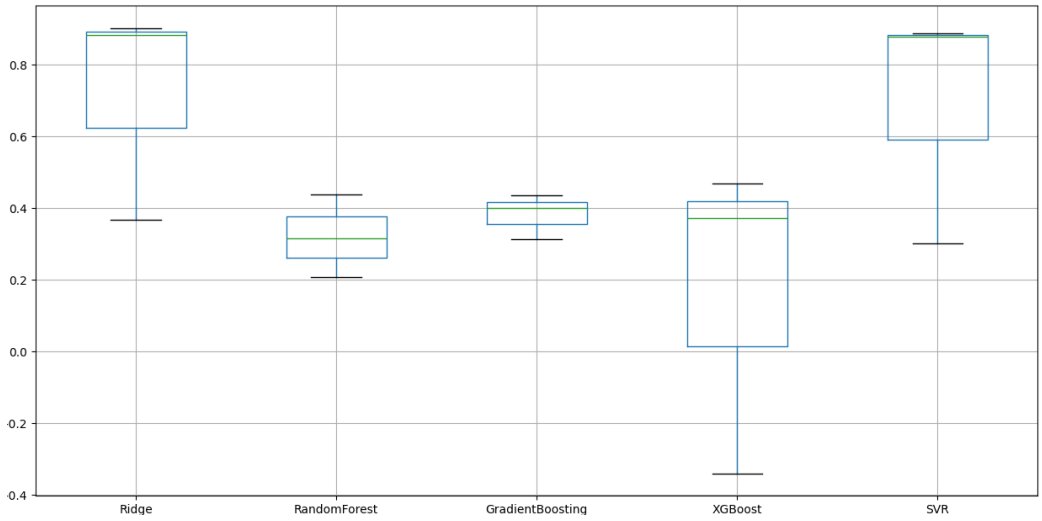

Figura 18: BoxPlot dei risultati degli algoritmi a confronto

Il *Box* (quindi il rettangolo) generalmente rappresenta l'intervallo che va dal primo quartile (Q1) al terzo quartile (Q3), dove il primo quartile è il punto sotto cui cade il 25% dei dati, mentre il terzo quartile è il punto sotto cui cade il 75% dei dati. In questo caso "l'altezza" del box rappresenta la dispersione degli R^2 ottenuti dalla *cross validation*. Quindi, un box più "basso" indica che i punteggi sono meno dispersi e quindi le previsioni del modello sono più consistenti. Un box più "alto" indica maggiore variabilità nelle previsioni.

La linea verde rappresenta la mediana, ovvero il valore centrale dei punteggi ottenuti durante la *cross-validation* (quindi il 50% dei punteggi cadrà sopra la linea e il 50% sotto). Quindi, una mediana alta indica che la maggior parte delle previsioni fatte dal modello ha prestazioni elevate.

I Baffi (le linee che si estendono dal *Box*) rappresentano l'intervallo dei punteggi che non sono considerati *outliers*. Indicano la gamma di punteggi "tipici" del modello. Potrebbero

anche comparire dei punti singoli esterni al rettangolo che rappresentano gli *outliers* (ovvero quei valori anomali che si discostano significativamente dalla distribuzione centrale). Generalmente la presenza di *outliers* indica che il modello ha avuto alcune previsioni con prestazioni molto diverse rispetto alla maggior parte delle previsioni. Pochi *outliers* possono essere normali, ma molti possono suggerire instabilità nel modello: fortunatamente in questo caso non ci sono.

Una volta chiarito il grafico è possibile trarre delle conclusioni: si escludono *Random Forest e XGBoost*. *Random Forest* a causa dell'R^2 (il più basso ottenuto), mentre *XGBoost* a causa della sua variabilità troppo grande (con anche valori negativi che sono indicatori di un modello pessimo, con prestazioni peggiori di una semplice media aritmetica). Verrà posto un focus solamente su regressione *Ridge, Gradient Boosting e SVR*: regressione *Ridge e SVR* grazie ai discreti risultati ottenuti, mentre *Gradient Boosting* grazie alla sua bassissima dispersione rispetto agli altri modelli.

5.8 Modello di ensemble

I risultati ottenuti in particolare con la regressione *Ridge e SVR* presi singolarmente sono discreti ma ancora migliorabili soprattutto in termini di dispersione (molto grande rispetto, ad esempio, del *Gradient Boosting*). Motivo per cui si proverà ad "unire" questi modelli per tentare di migliorare il risultato finale.

Per fare ciò verrà tentato un approccio di *ensemble*. Per ensemble si intende una tecnica di *machine learning* utile per combinare le previsioni di diversi modelli individuali per migliorare la performance predittiva rispetto all'uso di un singolo modello. L'idea è che combinando modelli diversi, portino una predizione più robusta e accurata.

Quindi, analogamente a quanto fatto precedentemente, viene caricato il dataset e pre-elaborato; successivamente, i diversi modelli di previsione vengono definiti, ciascuno con i propri parametri ottimali (ottenuti precedentemente tramite i test fatti con i singoli algoritmi). Il modello viene quindi valutato utilizzando *la cross-validation*: l'R^2 medio ottenuto da questo test è stato circa di 0.70, mentre l'RMSE di 26, indicandoci un modello di previsione decisamente buono vista l'alta variabilità dei dati.

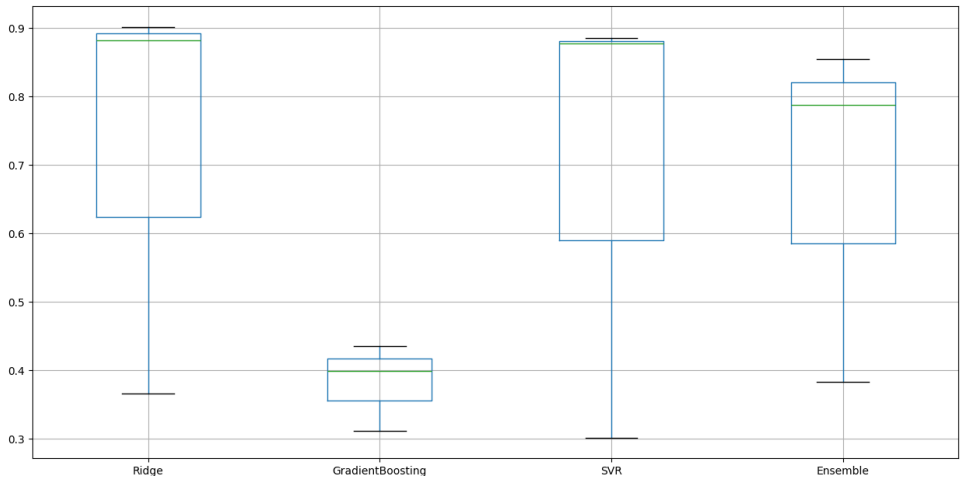

Figura 19: Confronto BoxPlot con Ensemble

Si evince come il modello di ensemble sia diventato la combinazione dei tre modelli: seppur sia leggermente peggiorato rispetto, ad esempio, al modello *Ridge* preso singolarmente, riesce a giovare della più bassa dispersione del *Gradient Boosting* (box e baffi molto più stretti rispetto agli altri due modelli).

5.9 Considerazioni

Il modello, nella forma in cui si trova attualmente è utile per fare previsioni del PUN per il giorno successivo.

È importante sottolineare però, che il PUN, come già spiegato precedentemente, è stabilito nel Mercato del Giorno Prima (MGP), questo significa che ogni giorno viene stabilito il PUN del giorno successivo, ampliando il periodo di previsione. Allo stesso modo funzionano molte variabili inserite nel dataset come il Punto di Scambio Virtuale o le materie prime.

È anche doveroso far notare che a differenza di quanto fatto in questo progetto, in cui i dati utilizzati sono giornalieri, è anche possibile sfruttare dati mensili o annuali, coprendo un periodo storico di gran lunga maggiore, così da poter avere anche stime approssimative del PUN del medio e del lungo periodo.

Anche parlando di dati utilizzabili si fa notare che quelli utilizzati sono dati pubblici disponibili per tutti. Nel caso in cui ci fosse stata la possibilità di sfruttare dati professionali, si potrebbero fare delle considerazioni su eventuali dati da aggiungere, come ad esempio quelli sul rapporto domanda/offerta. Oppure anche su dati da sostituire perché o troppo difficili da reperire o privati. Un esempio può essere quello dei dati

sull'indice di rischio geopolitico: quelli utilizzati sembrerebbero dati molto validi e ben fatti, però non sono comunque derivati da una fonte "attendibile" come può essere "BlackRock" che si occupa, tra le altre cose, della creazione di questa tipologia di indici.

5.9.1 Possibili implementazioni

Quello appena spiegato non è altro che la base per creare un modello di previsione molto più efficiente ed effettivamente utile nella vita reale. Il modello, infatti, si basa esclusivamente su dati storici senza mostrarne un effettivo utilizzo "pratico".

Si possono implementare anche delle "API" per avere dati più dettagliati in tempo reale; così da avere costantemente previsioni per il futuro prossimo: che sia di qualche ora o fino a due giorni avanti. Un "API", ossia *"Application Programming Interface"*, è un meccanismo che stabilisce il modo in due software differenti possono comunicare tra loro. Nello specifico, può servire per stabilire il metodo in cui un software può richiedere servizi e dati che provengono da un altro software in maniera autonoma. Questo modello di previsione, per esempio, può essere utile per alcune aziende per la compravendita di energia elettrica, con l'obiettivo di massimizzare il loro risparmio energetico, prevedendo picchi e cali del PUN prima che si verifichino.

Oppure può essere utile per la previsione della produzione di elettricità. Se, ad esempio, un'azienda produce energia, le predizioni del PUN possono essere utilizzate per determinare il momento in cui devono aumentare o ridurre la produzione per avere più guadagno, o costi inferiori delle materie prime.

Oppure, ancora, può essere utile per le aziende che stipulano contratti di fornitura di energia. Queste possono utilizzare le previsioni del PUN per negoziare migliori condizioni contrattuali. Per farlo si potrebbero utilizzare le previsioni per identificare i periodi in cui il PUN è previsto essere basso e stipulare contratti di fornitura durante questi periodi.

6 CONCLUSIONE

In questa tesi, è stato presentato un modello predittivo per il Prezzo Unico Nazionale nel mercato italiano utilizzando diverse tecniche di *machine learning* implementate per tenere conto delle dinamiche del mercato italiano, incluse perturbazioni quali la pandemia di COVID-19 e la guerra in Ucraina. I risultati ottenuti dimostrano che anche con la complessità e variabilità dei dati, la combinazione delle diverse tecniche di *machine learning* utilizzate garantiva previsioni relativamente precise.

In particolare, il modello di *ensemble* ha portato a una riduzione della varianza del PUN più efficace rispetto agli altri algoritmi presi singolarmente, incrementando la stabilità e la flessibilità delle previsioni. Nel corso del lavoro, sono stati analizzati i vari fattori che influenzano i prezzi dell'energia elettrica, tra cui la domanda, le materie prime e l'incertezza geopolitica. Il più impattante e correlato risulta essere il gas naturale. Spiccano anche il petrolio e il carbone per loro rilevanza.

Un problema critico è la volatilità dei dati, che spiega la limitata flessibilità e stabilità del modello. La regressione *Ridge* e il *Support Vector Regression* sono stati scelti per mitigare tale limite, e il loro effetto sul RMSE e R^2 è stato efficace. Potenziali sviluppi futuri includono l'integrazione di dati in tempo reale tramite API, che possono portare una maggior efficacia per le previsioni a breve termine. Inoltre, sarebbe interessante esplorare l'utilizzo di dati più specifici e dettagliati, come il rapporto domanda/offerta e altre variabili macroeconomiche, che potrebbero contribuire a migliorare ulteriormente la qualità delle previsioni.

7 BIBLIOGRAFIA E SITOGRAFIA

7.1 Bibliografia

[1] Decreto Legislativo n. 79/1999, noto come Decreto Bersani.

[2] PNRR (Piano Nazionale di Ripresa e Resilienza) - Sezione "Rivoluzione verde e transizione ecologica".

[3] Sorgenia.it – Informazioni base sul mercato energetico ed il PUN. Disponibile su: sorgenia.it.

[4] Breiman, L. (2001). Random Forests. *Machine Learning*.

[5] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

[7] Vapnik, V. (1995). The Nature of Statistical Learning Theory (SVR).

[8] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (selezione Iperparametri).

[9] James H. Stock, Mark W. Watson (2020). “Introduzione all’econometria”.

[10] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (SVR).

[11] Caldara & Iacoviello, M. Measuring Geopolitical Risk. Working Paper, Federal Reserve Board. Disponibile su: https://www.matteoiacoviello.com/gpr.htm.

[12] ERA5 – ECMWF Reanalysis: European Centre for Medium-Range Weather Forecasts. Guida per l’elaborazione dei dati del database climatico.

7.2 Sitografia

[12] GME (Gestore dei Mercati Energetici) - Dati storici sul Prezzo Unico Nazionale (PUN) e immagine introduzione. Disponibile su: mercatoelettrico.org.

[13] Investing.com - Dati storici delle materie prime come gas, petrolio e carbone e indice di emissione di CO_2. Disponibile su: investing.com.

[14] Snam Rete Gas - Dati storici relativi al Punto di Scambio Virtuale (PSV). Disponibile su: snam.it.

[15] Terna S.p.A. - Dati sul fabbisogno energetico italiano e immagine introduzione. Disponibile su: terna.it.

[16] Copernicus (European Centre for Medium-Range Weather Forecasts) - Dati climatici storici. Disponibile su: cds.climate.copernicus.eu.

[17] Our World in Data - Dati sulle fonti di generazione di energia elettrica in Italia. Disponibile su: ourworldindata.org.

[18] Google Maps - Utilizzato per la selezione delle aree nei dati meteorologici. Disponibile su: maps.google.com.

[19] Segugio.it - Dati giornalieri del Punto di Scambio Virtuale (PSV) forniti da Snam. Disponibile su: segugio.it.

[20] ResearchGate - Risorsa per la figura e la spiegazione della cross-validation rolling window. Disponibile su: researchgate.net.

[21] MathWorks - Fonte per la figura e la spiegazione della cross-validation "standard". Disponibile su: mathworks.com.

7.3 Lavori simili

[23] Meneghetti Sara (a.a. 2022/2023). Il mercato elettrico italiano. Modelli statistici di previsione dei prezzi e della domanda (Università degli Studi di Padova). Reperibile da: https://thesis.unipd.it/retrieve/631052b7-be2f-448a-a63a-22b5617427f9/Meneghetti_Sara.pdf

[24] Francesca Remigio (a.a. 2022/2023). I prezzi dell'energia elettrica nel Nord Italia: modelli di previsione (Università degli Studi di Padova). Reperibile da: https://thesis.unipd.it/retrieve/d48181ba-89bd-4a1e-bac3-f2cd1b7c4a26/Remigio_Francesca.pdf

[25] Axel Myrberger (2022). A machine learning approach for electricity future price prediction (KTH Royal Institute of Technology - Stockholm). Reperibile da: https://www.diva-portal.org/smash/get/diva2:1713311/FULLTEXT01.pdf