Analisi predittiva e interpretativa mercati finanziari: un modello basato sul *machine learning* e sulle tecniche di ottimizzazione

Abstract

Questo studio esplora l'applicazione di modelli di *machine learning* – la cui messa a punto è stata resa più efficiente utilizzando opportune tecniche di ottimizzazione – per interpretare i fattori più rilevanti nelle scelte degli investitori e prevedere i movimenti aggregati dei prezzi delle azioni nei mercati finanziari. Utilizzando dati storici, tra cui volumi, indici di volatilità e indicatori finanziari e macroeconomici, i modelli hanno raggiunto un'elevata accuratezza predittiva, evidenziata dai risultati conseguiti in termine di coefficiente di determinazione (R^2), per il quale è stata anche proposta una formulazione alternativa applicabile a serie limitate come quelle che hanno caratterizzato alcune gravi crisi finanziarie mondiali.

Angelo Anglani, Claudio Andrioli, Filippo Alessandro Fornaro

Indice

1. Introduzione

La caratterizzazione quantitativa dei fattori incidenti sulle scelte degli investitori sui mercati finanziari costituisce storicamente una sfida estremamente impegnativa ed affascinante. Sebbene vi siano correlazioni note ed ampiamente considerate dagli operatori nelle proprie scelte come il ruolo fondamentale dei tassi di interesse e dell'occupazione sulle dinamiche generali degli investimenti e dei prezzi[1], come anche quello della massa monetaria[2], le relazioni identificate sono perlopiù qualitative. È altresì riconosciuto come vi sia una rilevanza molto grande delle cosiddette "aspettative"[3] nel comportamento degli operatori economici e conseguentemente nella predizione delle relative scelte, dalle quali dipendono i valori presenti e futuri degli indicatori economici e finanziari. Allo stesso tempo è altrettanto evidente la difficoltà nel rappresentare quantitativamente in modo efficace tali "aspettative" e nel modellare gli andamenti delle grandezze oggetto di studio in dipendenza di tali rappresentazioni.[4] Tali considerazioni sono di fondamentale importanza nell'approccio all'interpretazione, e successivamente alla possibile predizione, degli andamenti dei mercati finanziari. In essi una componente fondamentale del prezzo dei titoli scambiati si basa sulla valutazione dell'investitore a proposito del guadagno atteso (e del rischio di perdita) nel lungo periodo, confrontato con la rinuncia al guadagno che si otterrebbe nel breve periodo mediante i convenzionali strumenti di remunerazione del capitale a breve termine.[5] Laddove la previsione che l'investitore ha fatto dovesse rivelarsi errata, peraltro, per i mercati finanziari la "*Speed of Expectations Revision*" e la "*Adjustment Dynamics*"[6] è pressoché istantanea, poiché i cambiamenti nelle strategie di allocazione degli investimenti e le reazioni alle variazioni dei parametri macroeconomici avvengono nel ridotto tempo necessario ad apprendere le notizie che confermano o smentiscono le predizioni – tra le quali, banalmente, vi è soprattutto il prezzo *mark-to-market* del titolo – e piazzare i conseguenti ordini di acquisto/vendita.

In base a quanto sopra esposto, pertanto, è possibile immaginare l'esistenza di un modello che, in un dato sistema economico, tenga conto quantitativamente:

- dei valori assunti dai parametri economici e macroeconomici che rappresentano, in qualche modo, il valore "sostanziale" degli *asset* che di tale sistema economico sono espressione;

---

[1] Si vedano i principi generali espressi da Sir John Maynard Keynes, "*The General Theory of Employment, Interest and Money*.", ovviamente, ma anche da Albert Ando, "*An Empirical Model of United States Economic Growth: An Exploratory Study in Applied Capital Theory*" nell'ambito del lavoro "*Models of Income Determination*" e da Ignazio Visco "*Dalla Teoria Alla Pratica Nei Modelli Macroeconomici: L'Eclettismo Post-Keynesiano*" da "*Moneta e Credito*".

[2] Il rapporto tra offerta di moneta e livelli dei prezzi è la colonna portante della corrente monetarista generalmente ricondotta all'opera di Milton Friedman ("*Monetary History of the United States 1867-1960*").

[3] Si veda Flint Brayton et al., "*The Role of Expectations in the FRB/US Macroeconomic Model*" sul ruolo attribuito alle aspettative nei modelli classici come MPS e nei nuovi modelli come il FRB.

[4] Flint Brayton et al.: "*The Role of Expectations in the FRB/US Macroeconomic Model*", testualmente: "*Economists have long recognized that expectations play a prominent role in economic decisionmaking and are a critical feature of macroeconomic models. However, they disagree about the basis on which individuals form expectations and thus about the way to model them*."

[5] Sempre in Flint Brayton et al.: "*Tying the current price of an asset to its expected future earnings is a common way of modeling bond and equity prices and is not unique to FRB/US…… Thus, expectations about future dividends, future inflation, and future short-term interest rates, as captured by the corporate bond rate, determine the current price of equity.*"

[6] Ancora in Flint Brayton et al.: "*In the FRB/US model, expectations about future economic conditions influence current prices and activity by means of two distinct channels. Through the first channel, asset valuation, today's price of an asset is linked to the expected earnings stream of the asset and the expected rate of return on alternative assets. Thus, in the model, current bond and stock prices are determined by the present discounted value of expected coupon and dividend payments. Through the second channel, adjustment dynamics, expectations play a role in reducing the costs of economic frictions.*"

- dell'espressione delle "aspettative" degli operatori, da ricavarsi in base ai valori ed alle variazioni dei parametri finanziari ed operativi del mercato su cui i suddetti *asset* sono scambiati, ma in base all'indice dei livelli generali dei prezzi;

e li utilizzi per stimare il valore aggregato dei prezzi degli *asset* stessi.

Vista la grande quantità di dati da considerare e la difficoltà di elaborare modelli analitici si è scelto di generare il modello tramite *machine learning*, per tentare successivamente di migliorarlo utilizzando tecniche di ottimizzazione.

2. Scopo

L'obiettivo che ci si è proposti è l'elaborazione di un modello di *machine learning* in grado di prevedere il valore giornaliero dell'indice azionario *Dow Jones Industrial Average* (DJIA) interpretandone le relazioni con i parametri finanziari e macroeconomici del sistema economico di riferimento (quello statunitense). Metodologia

Sono stati addestrati modelli di regressione lineare e SVR (*Support Vector Regression*) in diversi intervalli temporali, inclusi eventi economici significativi come la crisi finanziaria del 2008 e la pandemia di COVID-19; i modelli sono stati quindi paragonati tra loro. Una volta identificato il modello più funzionale allo scopo che, volta per volta, ci si è prefissi, esso è stato ottimizzato allo scopo di ridurre l'insieme delle variabili indipendenti considerate, eliminando quelle molto correlate tra di loro e lasciando solo quelle con maggiore impatto sulla variabilità della variabile dipendente. Quest'ultimo processo è stato inizialmente condotto considerando le correlazioni mutue tra le varie variabili indipendenti ed ottimizzando il numero di variabili alla ricerca del miglior compromesso tra accuratezza e complessità del modello, utilizzando il software CPLEX per eliminare le variabili che presentavano una correlazione mutua superiore ad una certa soglia, imponendo la minimizzazione della funzione obiettivo esprimente il MSE (*Mean Square Error*). Un ulteriore tentativo di individuare le variabili "core" è stato in un secondo momento implementato mediante un ulteriore modello di ottimizzazione, ottenuto impostando come funzione obiettivo da massimizzare quella esprimente il valore di R^2 *adjusted*.

Ciò ha consentito di identificare un set di variabili indipendenti "*core*" da considerare e da includere nel modello da addestrare, alle quali eventualmente affiancarne altre per migliorare le capacità interpretative e/o predittive in contesti particolari.

2.1. Descrizione dei Dati

Il *dataset* include prezzi storici delle azioni e indicatori economici dal 1979 al 2024. Tra di essi, in particolare:

- **Indici, prezzi e volumi di scambio**:

  - **DJIA - Price** (fonte: sito www.investing.com[7]);
  - **DJIA - Open** (fonte: sito www.investing.com[7]);
  - **DJIA - High** (fonte: sito www.investing.com[7]);
  - **DJIA - Low** (fonte: sito www.investing.com[7]);
  - **DJIA - Vol.** (fonte: sito www.investing.com[7]);
  - **DJIA - Change %** (fonte: sito www.investing.com[7]);
  - **S&P 500 - P/E[8] Ratio** (fonte: sito www.macrotrends.net[9]);

[7] Consultato in data 14/12/2024.
[8] *Price/Earning*, ossia rapporto tra prezzo del titolo azionario e profitti da esso generati.
[9] Consultato in data 03/11/2024.

- **Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis** (fonte: *Board of Governors of the Federal Reserve System (US)*);
- **Market Yield on U.S. Treasury Securities at 20-Year Constant Maturity, Quoted on an Investment Basis** (fonte: *Board of Governors of the Federal Reserve System (US)*).

- **Indici di Volatilità**: Misure di incertezza del mercato, come il CBOE *Volatility Index* (VIX):

  - **CBOE Volatility Index: VIX** (fonte: *Chicago Board Options Exchange*)**;**
  - **CBOE Volatility Index: VIX_1st_derivative** (variazione marginale del parametro VIX, calcolato a partire dalle istanze giornaliere dello stesso);

- **Indicatori Economici**:

  - **Real Gross Domestic Product** (fonte: *U.S. Bureau of Economic Analysis*);
  - **Federal Surplus or Deficit [+/-]** (fonte: *U.S. Office of Management and Budget*);
  - **St. Louis Fed Financial Stress Index** (fonte: *Federal Reserve Bank of St. Louis)*;
  - **Unemployment Rate** (fonte: *U.S. Bureau of Labor Statistics*);
  - **Federal Debt: Total Public Debt** (fonte: *U.S. Department of the Treasury, Fiscal Service*);
  - **Federal Funds Effective Rate** (fonte: *Board of Governors of the Federal Reserve System (US)*);
  - **M2** (fonte: *Board of Governors of the Federal Reserve System (US)*);
  - **Consumer Price Index for All Urban Consumers: All Items in U.S. City Average** (fonte: U.*S. Bureau of Labor Statistics*).

Tutte le suddette serie storiche sono state considerate, all'inizio, come possibili variabili indipendenti (X) del modello, tranne il "*DJIA – Price*" (di seguito indicato genericamente come "*Price*" o "prezzi") che è stato identificato come variabile dipendente (Y). Il *St. Louis Fed Financial Stress Index* ha mostrato fin dai primi test preliminari una bassa correlazione con la variabile dipendente. Poiché le relative istanze sono disponibili solo a partire dall'inizio degli anni '90, rendendo non utilizzabile quasi dieci anni di dati (tutti gli altri attributi sono disponibili contemporaneamente a partire da fine anni '70), si è deciso di procedere senza utilizzare tale *feature*. Anche gli attributi direttamente collegati al "*Price*", come "*DJIA – Open*", "*DJIA – High*" e "*DJIA - Low*" sono stati inizialmente stralciati, per evitare di inserire tra le variabili indipendenti versioni "camuffate" della variabile dipendente togliendo significatività al modello.

Alcuni dei parametri sopra menzionati sono riferiti all'indice S&P500 (*Standard and Poor* 500), che è un altro indice basato sul mercato azionario statunitense, calcolato in maniera lievemente differente. Essendo tuttavia il sottostante (ossia l'insieme di titoli del cui valore i due indici costituiscono una rappresentazione sintetica), di fatto, il medesimo, i due indici sono estremamente correlati e costituiscono entrambi una valida ed affidabile rappresentazione aggregata dei prezzi dei titoli azionari statunitensi.

2.2. *Preprocessing*

Le variabili indipendenti presentano una cadenza di campionamento non omogenea (annuale, trimestrale, settimanale, giornaliera). Il primo passaggio per poterli rendere utilizzabili è stato quindi quello di omogeneizzarne la distribuzione, replicando con granularità giornaliera tutte quelle grandezze aventi distribuzione settimanale, trimestrale o annuale. Il *dataset* così ottenuto, contenuto in un file *Excel*, è stato quindi avviato alla *pipeline* di *preprocessing* dinamica effettuata mediante la funzione *download_data* contenuta nello script Python che implementa il modello. In tale funzione i successivi passaggi di *preprocessing* dei dati prevedono:

a) Lettura dei dati

I dati vengono letti da un file *Excel* specificato in una variabile di controllo e caricati in un *DataFrame* Pandas.

b) Pulizia iniziale dei dati

Tutti i dati inconsistenti e/o inutilizzabili vengono marcati come NaN (*Not-a-Number*) e quindi eliminati insieme alle righe del *DataFrame* in cui sono presenti.

c) Gestione della colonna "*Price*"

È possibile scegliere se considerare "*Price*" esclusivamente come variabile dipendente, oppure includerne le relative istanze precedenti (al proposito si veda il seguente punto "e)") tra le variabili indipendenti, o addirittura eliminare completamente la colonna dal *dataset*.

d) Aggiunta di variabili indipendenti

Se richiesto, vengono aggiunte le seguenti *features*:

- "*MME_Prices*": medie mobili esponenziali calcolate sulla colonna "*Price*".
- "*MME_vs_Prices*": differenza ponderata (tramite il parametro "*MMEPricesWeightModel*") tra il valore del prezzo e la media mobile.
- "*PriceVariation*" variazione marginale dei prezzi, calcolata a partire dalle istanze giornaliere degli stessi.

e) Creazione di istanze precedenti delle variabili indipendenti

In base a quanto espresso nei parametri di configurazione, possono essere incluse delle colonne basate sulle ultime k-1 osservazioni per le seguenti *features*:

- "*DJIA – Price*" (colonne "*Price-1*", "*Price-2*", ..., "*Price-(k-1)*");
- "*DJIA - Vol.*";
- "*CBOE Volatility Index: VIX*" e "*CBOE Volatility Index: VIX_1st_derivative*";
- "*Change %*";
- "*PriceVariation*".

f) Identificazione del *target*

La colonna contenente le istanze della variabile dipendente prescelta vengono posizionate alla destra di tutte le altre variabili; in base all'orizzonte di predizione f le istanze vengono traslate di f posizioni per far corrispondere i valori delle variabili indipendenti al giorno i con il valore assunto dalla variabile dipendente al giorno i+f.

g) Restituzione del *DataFrame* con tutte le trasformazioni applicate

h) Salvataggio dei dati preprocessati

Il *DataFrame* preprocessato viene salvato in un foglio ("*DatasetFormattatoNoNaN*") di un nuovo file *Excel* specificato dall'utente contenente gli *output*.

Al *DataFrame* così approntato, a scelta dell'utente, possono essere applicate ulteriori trasformazioni di standardizzazione (mediante *MinMaxScaler*), di normalizzazione (mediante *StandardScaler*), oppure entrambe, a scelta dell'utente.

2.3. Modelli

Il *dataset* è stato suddiviso in due parti (*train* e *test*) composti ciascuno dal 50% dei campioni disponibili, su base casuale. Il *dataset* di *train* è stato quindi utilizzato per addestrare i seguenti modelli:

1. **Regressione Lineare**: il modello base per stabilire una relazione lineare tra variabili indipendenti e prezzi delle azioni. Si tratta di un approccio bilanciato per valutare ed interpretare in modo diretto l'impatto di ciascuna variabile indipendente su quella dipendente.

2. **Support Vector Regression (SVR)**: sono stati testati modelli non lineari con *kernel* a funzione radiale (RBF), lineari e polinomiali. Le caratteristiche di non-linearità potrebbero meglio approssimare le situazioni in cui gli indici hanno subito forti stress dovuti a cause esogene, come in alcuni degli intervalli temporali evidenziati al successivo paragrafo. A parte la dichiarazione del *kernel* da utilizzare volta per volta, per questi modelli sono stati utilizzati i parametri di *default*. Fanno eccezione il *gamma* per il modello RBF, specificato pari a 0.1 come compromesso tra granularità e generalità, e il *degree* per il modello polinomiale, impostato a 2 per non enfatizzare eccessivamente la componente non lineare.

3. **Random Forest**: dopo le prime simulazioni, una volta condotta una valutazione preliminare sui modelli sopra descritti – più facilmente interpretabili – come terzo approccio è stato implementato un ulteriore modello basato su un insieme di alberi decisionali, le cui previsioni vengono combinate per fornire l'*output* finale. Questo metodo è particolarmente robusto nei confronti di dati rumorosi e relazioni non lineari, poiché utilizza il principio della media (in questo caso, in cui si sta implementando un modello di regressione su una variabile continua, mentre quando si trattano variabili categoriche si utilizza generalmente il voto a maggioranza) per ridurre il rischio di *overfitting* associato ai singoli alberi. Inoltre la *random forest* permette di valutare l'importanza relativa delle variabili indipendenti, fornendo utili considerazioni sui fattori che influenzano maggiormente i prezzi delle azioni.

I modelli sono stati successivamente testati sia sul *dataset* di *test* che sull'intero *dataset* originario allo scopo di fornire un riscontro grafico immediato dell'accuratezza o meno dei modelli generati, andando poi in particolare ad analizzare le *performance* conseguite in alcuni periodi temporali molto particolari, corrispondenti alle principali crisi finanziarie degli ultimi 30 anni, di seguito descritti nel dettaglio.

2.4. Intervalli Temporali e Metriche

L'analisi dei modelli è stata condotta sull'intero *dataset* disponibile, che copre un periodo esteso dal **25 dicembre 1979** al **1° novembre 2024**, e su specifici intervalli temporali di particolare interesse storico ed economico. Gli intervalli selezionati includono:

- **Bolla Dot-com**: 1 gennaio 2000 – 31 dicembre 2001;
- **Attacchi dell'11 Settembre**: 1 aprile 2001 – 30 aprile 2002;
- **Crisi Finanziaria del 2008**: 1 luglio 2008 – 31 luglio 2009;
- **Crisi del Debito Sovrano Europeo**: 1 aprile 2011 – 30 aprile 2012;
- **Pandemia di COVID-19**: 1 ottobre 2019 – 30 ottobre 2020.

Questi periodi sono stati scelti per testare la robustezza dei modelli in condizioni di mercato volatili e in contesti economici eccezionali, poiché rappresentano fasi caratterizzate da dinamiche economiche e finanziarie imprevedibili, eventi globali di grande impatto e rapidi cambiamenti nei mercati. Analizzare le prestazioni dei modelli in questi intervalli consente di valutare la loro capacità di adattarsi a condizioni estreme, come improvvise crisi di liquidità, drastici cali dei mercati azionari, turbolenze economiche causate da eventi geopolitici o sanitari e cambiamenti strutturali nei comportamenti degli investitori. Questa analisi

aiuta a identificare i limiti dei modelli e la loro affidabilità nel fornire previsioni accurate in scenari reali che esulano dalle normali condizioni di mercato.

2.4.1. Metriche di Valutazione

La valutazione delle prestazioni dei modelli è stata effettuata utilizzando le seguenti metriche standard del *machine learning* e della regressione:

1. **Mean Squared Error (MSE)**: misura l'errore medio quadratico tra i valori osservati e quelli predetti. Valori più bassi indicano una migliore accuratezza del modello.
2. **Mean Absolute Error (MAE)**: calcola l'errore medio assoluto, ossia la differenza media in valore assoluto tra i valori reali e quelli previsti.
3. **Root Mean Squared Error (RMSE)**: radice quadrata del MSE, rappresenta un'indicazione chiara dell'errore medio in unità della variabile *target*.
4. **R^2 (coefficiente di determinazione)**: valuta la proporzione della varianza dei dati spiegata dal modello. Un valore vicino a 1 indica una forte capacità predittiva.

Ad esse è stata aggiunta la seguente metrica, non convenzionale e definita in occasione del presente studio:

5. **R^2 parziale con varianza globale (R^2 ridotto)**: calcola un R^2 modificato che confronta la varianza dei valori in un periodo temporale specifico rispetto alla varianza dell'intero *dataset*. Questa metrica fornisce una misura della capacità del modello di mantenere la coerenza predittiva in contesti temporali limitati, tenendo conto della variabilità complessiva dei dati. L'esigenza di disporre di tale metrica, la sua formulazione le sue caratteristiche sono oggetto del successivo paragrafo 3.4.2.

Le metriche sono state calcolate sia sull'intero *dataset* sia su ciascun intervallo temporale definito, consentendo un confronto dettagliato delle prestazioni dei modelli in condizioni di mercato diverse.

2.4.2. Osservazioni sulle metriche: una formulazione alternativa per R^2

L'analisi delle *performance* dei modelli nei periodi ridotti corrispondenti alle crisi finanziarie ha evidenziato alcune peculiarità singolari della metrica R^2. La formulazione classica di R^2 è la seguente:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{ESS_r}{TSS_r}$$

con RSS devianza residua (*Residual Sum of Squares*):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

TSS devianza totale (*Total Sum of Squares*):

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

ed infine ESS (*Explained Sum of Squares*) la devianza spiegata dal modello:

$$ESS = \sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Nell'andare tuttavia a calcolare R^2 su intervalli ridotti rispetto a all'intervallo completo su cui si è testato il modello, il valore del TSS viene fortemente a dipendere da $\overline{y}$ riferito all'intervallo ridotto. In tal modo vi è un effetto distorsivo legato al fatto che la devianza residua viene confrontata con la devianza assunta dalla y esclusivamente nell'intervallo ridotto, e non con la devianza totale; tale effetto può divenire considerevole nel caso in cui la devianza della y sia molto ridotta nell'intervallo considerato.

Allo scopo pertanto di confrontare omogeneamente la *performance* conseguita nell'intervallo ridotto rispetto a quella dell'intero *dataset* si è ipotizzata (ed utilizzata) una formulazione alternativa per la metrica, ottenuta sostituendo a $\overline{y}$ calcolato nell'intervallo ridotto con $\overline{y}$ esteso all'intero intervallo di *test.*

Definiti a questo punto:

$$R_r{}^2 = 1 - \frac{RSS_r}{TSS_r} = 1 - \frac{\sum_{i=1}^{n}(y_{ri} - \hat{y}_{ri})^2}{\sum_{i=1}^{n}(y_{ri} - \overline{y_r})^2} = \frac{ESS_r}{TSS_r}$$

il R^2 secondo la formulazione convenzionale applicato all'intervallo ridotto, caratterizzato da $\overline{y_r}$ intesa come media dei valori assunti (denominati y_{ri}) da y nell'intervallo ridotto, le cui stime sono identificate con $\hat{y}_{ri}$ e:

$$R_t{}^2 = 1 - \frac{RSS_t}{TSS_t} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_t})^2} = \frac{ESS_t}{TSS_t}$$

il R^2 secondo la formulazione convenzionale applicato all'intero intervallo di test, caratterizzato da $\overline{y_t}$ intesa come media dei valori assunti da y nell'intervallo totale, possiamo infine definire il **Coefficiente di Determinazione Ridotto (**R_{rt}^2**)** pari a:

$$R_{rt}{}^2 = 1 - \frac{RSS_r}{TSS_{rt}} = 1 - \frac{\sum_{i=1}^{n}(y_{ri} - \hat{y}_{ri})^2}{\sum_{i=1}^{n}(y_{ri} - \overline{y_t})^2} = \frac{ESS_r}{TSS_{rt}}$$

che caratterizza la quota di devianza spiegata dal modello nell'intervallo ridotto rispetto alla quota di devianza assunta nell'intervallo ridotto dalla Y rispetto alla media della Y nell'intervallo globale.

Un valore di R_{rt}^2 maggiore di $R_t{}^2$ indica che il modello spiega la devianza nell'intervallo meglio di quanto faccia nell'intero intervallo di test; un valore inferiore rappresenta il contrario.

E' anche possibile definire il **Coefficiente di Determinazione Relativo (**R_l^2**)** pari a:

$$R_l^2 = \frac{R_{rt}^2}{R_t^2} = \frac{1 - \frac{\sum_{i=1}^{n}(y_{ri} - \hat{y_{ri}})^2}{\sum_{i=1}^{n}(y_{ri} - \overline{y_t})^2}}{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_t})^2}} = \frac{ESS_r}{TSS_{rt}} \cdot \frac{TSS_t}{ESS_t} = \frac{ESS_r}{ESS_t} \cdot \frac{TSS_t}{TSS_{rt}} = \frac{ESS_l}{TSS_l}$$

Il quale misura la prevalenza tra la quota di devianza spiegata dal modello nell'intervallo rapportata alla devianza spiegata nel totale (ossia ESS_l) e la devianza nell'intervallo rapportata alla devianza totale.

Analogamente a R_{rt}^2 un valore di R_l^2maggiore di 1 indica che il modello spiega la devianza nell'intervallo meglio di quanto faccia nell'intero intervallo di test; un valore inferiore a 1 rappresenta il contrario.

3. Simulazione iniziale: risultati

Sono stati implementati ed addestrati inizialmente 4 modelli, tre dei quali basati su SVR (*kernel* RBF, lineare e polinomiale), mentre l'ultimo modello è un modello di regressione lineare multivariato.

Operando sul *dataset* iniziale in modo da realizzare:

- l'aggiunta delle ultime 7 osservazioni per gli attributi:
    - "*Vol.*";
    - "*CBOE Volatility Index: VIX*";
    - "*CBOE Volatility Index: VIX_1st_derivative*";
- la normalizzazione dei dati;
- l'impostazione come variabile dipendente di "*Price*";
- 10 *run* di addestramento su 10 differenti *dataset* di *train* randomizzati sul 50% dei campioni;

si ottengono i seguenti risultati:

RUN #	RBF Model_RMSE	RBF Model_MAE	RBF Model_R2	Linear Model_RMSE	Linear Model_MAE	Linear Model_R2	Polynomial Model_RMSE	Polynomial Model_MAE	Polynomial Model_R2	Linear Regression_RMSE	Linear Regression_MAE	Linear Regression_R2
0	0,051630532	0,045402308	0,956915373	0,047338037	0,040342553	0,96378157	0,055045251	0,048027225	0,951027886	0,032808286	0,024902726	0,982602926
1	0,051876654	0,045321222	0,956648565	0,049015956	0,042731208	0,961297901	0,055129191	0,047775476	0,951042099	0,032737758	0,024927674	0,982735364
2	0,051417291	0,045318265	0,95709166	0,047676065	0,040513442	0,963108685	0,055299861	0,04850212	0,9503669	0,033235096	0,024773093	0,982072605
3	0,051140367	0,044986574	0,958936772	0,046997249	0,040013644	0,965320703	0,05583745	0,048770723	0,951047309	0,033577273	0,025476503	0,982298247
4	0,05176079	0,045131106	0,956984826	0,049459282	0,043154028	0,960725062	0,055348516	0,04819875	0,950815092	0,032401912	0,024701279	0,983143742
5	0,051170862	0,044670477	0,958105454	0,046873995	0,03981028	0,964845903	0,054996581	0,047843088	0,951606904	0,033404528	0,025063479	0,982146527
6	0,051579275	0,045185362	0,957859056	0,047622736	0,040528234	0,964076183	0,055325394	0,048361739	0,95151551	0,033648531	0,025393109	0,982065635
7	0,051797725	0,045108308	0,956865833	0,049985046	0,043702902	0,959831998	0,05485864	0,047587182	0,951617298	0,033308964	0,025217866	0,982162981
8	0,051499861	0,045199463	0,957493245	0,047746116	0,04072436	0,963463922	0,054881816	0,048065384	0,951727168	0,032552174	0,024722435	0,983017335
9	0,051804288	0,045507361	0,956719491	0,047889849	0,041235856	0,963013105	0,0555819	0,048422894	0,950177247	0,032874618	0,024697782	0,982570603
Finale	0,051768033	0,045430929	0,956929303	0,049567015	0,043149789	0,960513912	0,055239319	0,04826566	0,950959464	0,0325872	0,024793168	0,982933178

Tabella 1: Risultati dei test sui vari modelli

i quali evidenziano come le *performance* del modello basato su regressione lineare siano globalmente le migliori fra quelli considerati. Anche i grafici relativi alla distribuzione dei residui confermano la suddetta valutazione, seppur anche in questo caso la distribuzione dei residui suggerisca la possibilità che vi siano delle dinamiche – di natura con ogni probabilità non lineare – non completamente "catturate" dal modello.

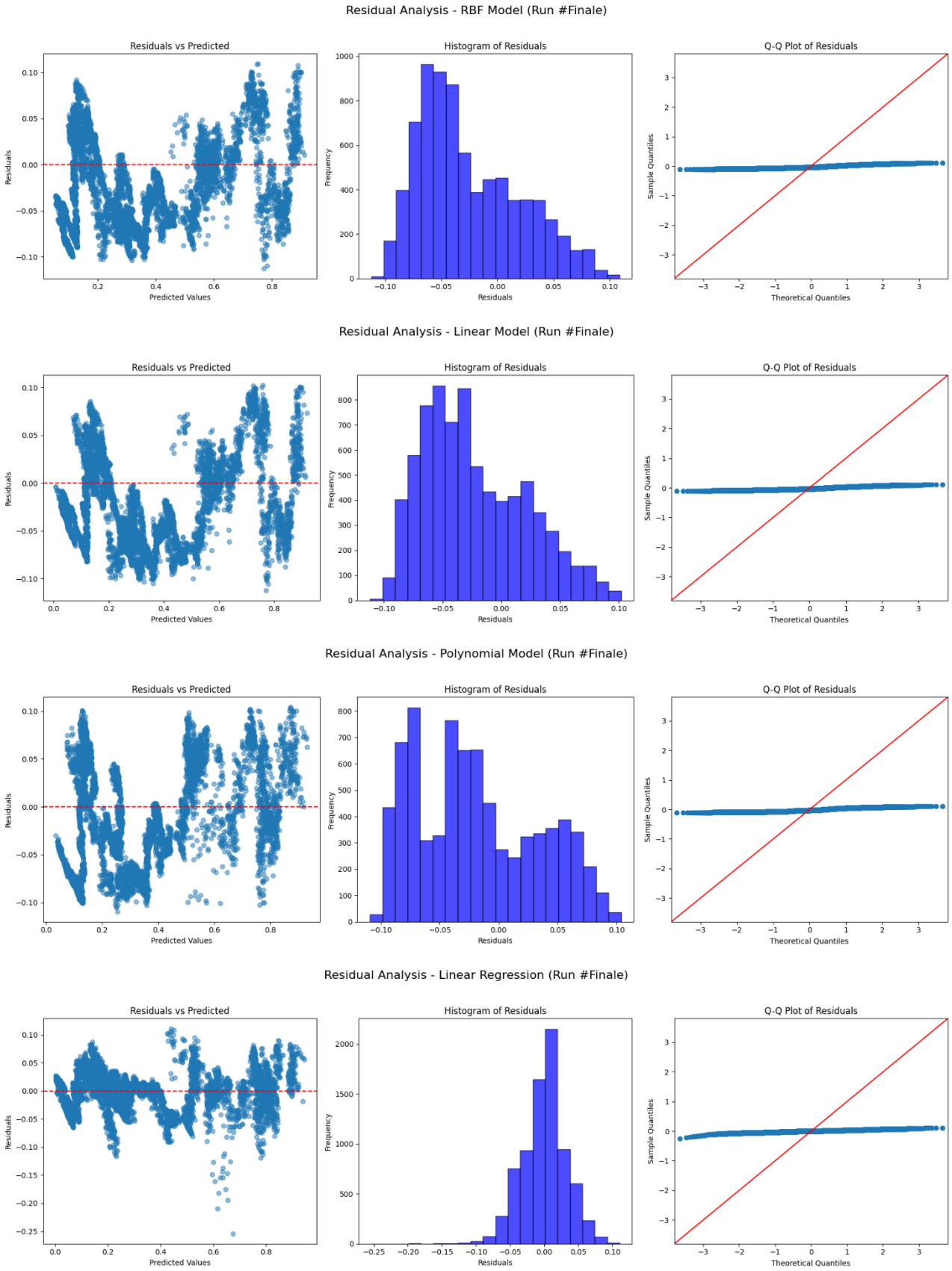


Figura 1: Distribuzioni dei residui

L'andamento generale predetto rispecchia abbastanza fedelmente quello realmente verificatosi.

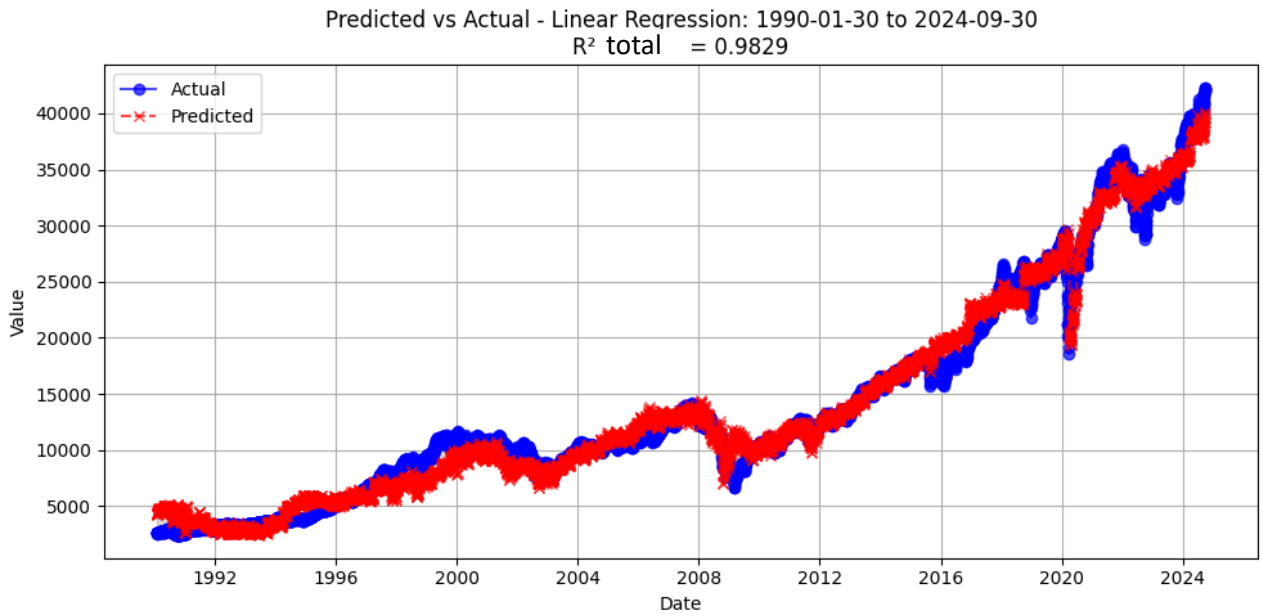


Figura 2: Confronto tra andamento predetto e andamento reale del DJIA

3.1. Considerazioni sull'accuratezza

Come si nota da quanto esposto sopra (Tabella 1), il modello basato sulla regressione lineare ha costantemente raggiunto valori di R^2 superiori a 0,982 in tutti i *run* effettuati. Proviamo ora ad analizzarne più in dettaglio il comportamento nei 5 intervalli temporali ridotti corrispondenti alle principali crisi finanziarie introdotte al paragrafo 3.4.

Ciò che si nota immediatamente è che la *performance* in ciascuno degli scenari considerati, seppur non bassa, è comunque sempre peggiore rispetto a quella globale: gli R^2_{rt} (R^2 ridotti) risultano sempre inferiori rispetto al ${R_t}^2$ (R^2 globale, pari a 0.9829).

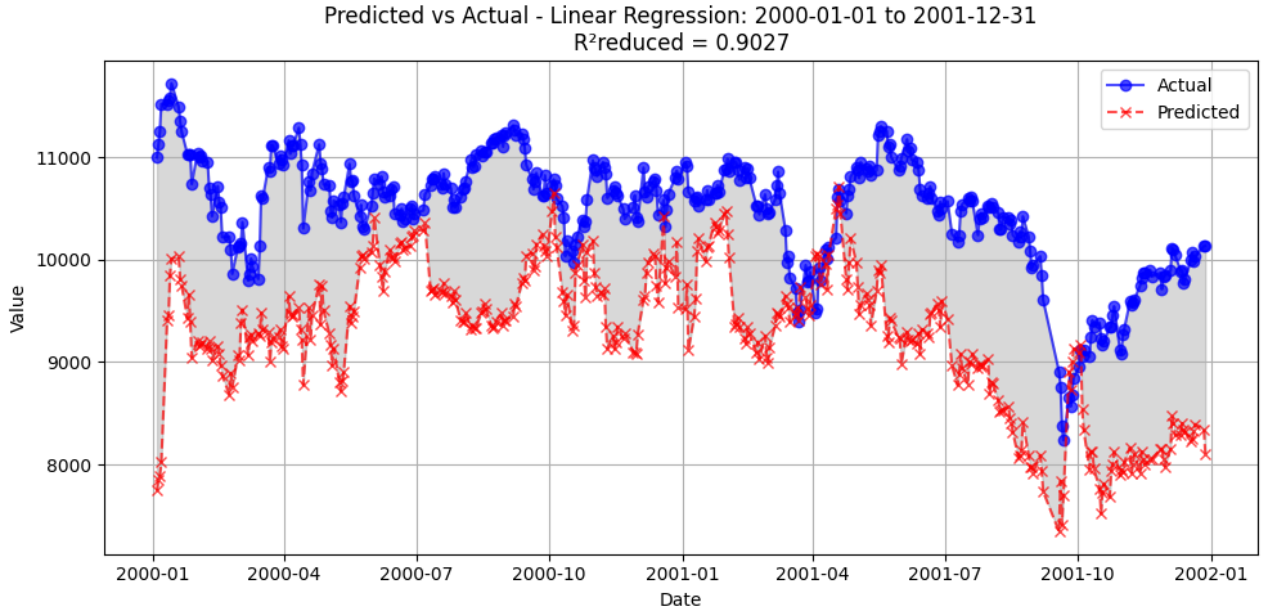


Figura 3*: Confronto tra andamento predetto e andamento reale DJIA – Bolla delle dot.com*

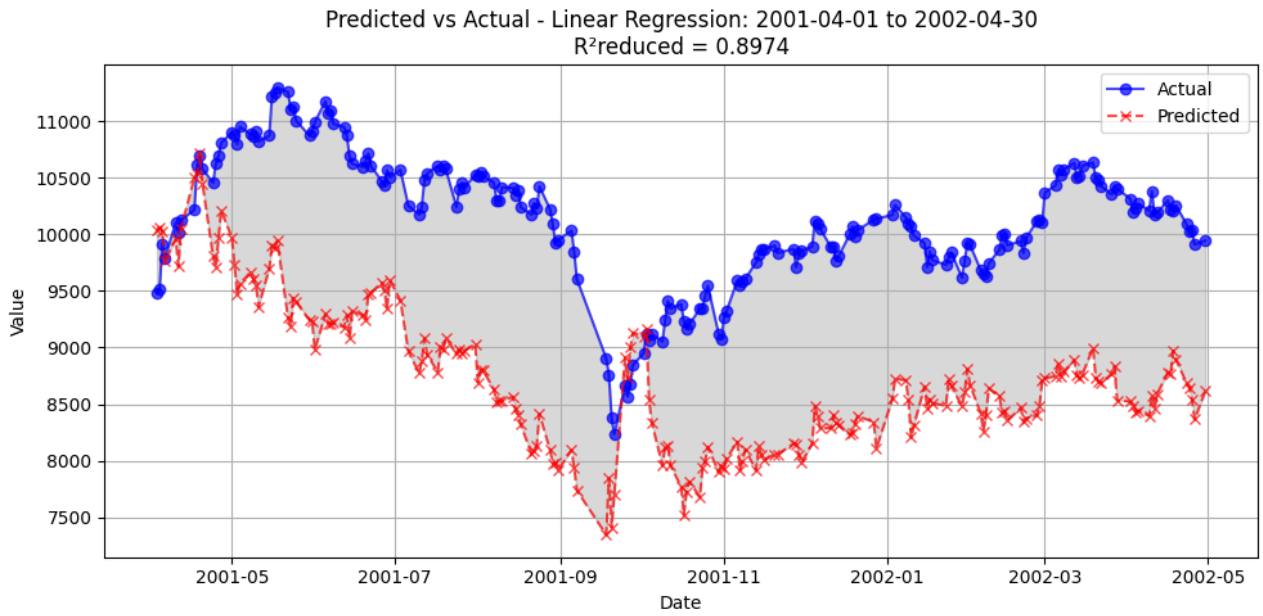


Figura 4: Confronto tra andamento predetto e andamento reale DJIA – Attacco alle Torri Gemelle (11 settembre)

Nei primi due scenari le previsioni del modello risultano conservative rispetto al fenomeno reale (Figura 3: – Bolla delle dot.com e Figura 4: Attacco alle Torri Gemelle), mostrando in ampi tratti un andamento abbastanza simile ma con un certo *bias*. Una possibile spiegazione è legata al sussistere di condizioni economiche e fondamentali non sufficienti (già a inizio 2000) a sostenere i livelli di prezzo assunti dalle azioni nel periodo, che quindi sono rapidamente crollati al manifestarsi di episodi negativi: prima, nel marzo 2000, la diffusione dei bilanci pubblicati da diverse aziende del comparto informatico e telematico – che avevano fino ad allora trainato la crescita degli indici – contenenti risultati e proiezioni deludenti, e poi, l'11 settembre 2001, l'attacco alle Torri Gemelle. Si noti come nella fase di crollo e immediatamente dopo gli andamenti predetti ed effettivi siano paralleli e si riconcilino per un breve tratto in modo perfetto, a suggerire che in tale lasso temporale prezzi e fondamentali (inclusivi chiaramente delle azioni correttive apportate dalla FED per stimolare la ripresa) si siano riallineati, salvo divergere successivamente per un nuovo disallineamento tra fondamentali e prezzi. Tale interpretazione si basa sulla considerazione del fatto che il modello di regressione lineare implementato trae le proprie predizioni da una combinazione lineare fra i parametri finanziari ed economici legati all'economia ed ai mercati statunitensi. Evoluzioni spiccatamente non lineari e livelli di prezzi non organicamente legati ai movimenti dei parametri fondamentali sono difficilmente catturati dal modello, che in tali situazioni mostra predizioni non accurate.

Tale interpretazione collima con la seguente Figura 5, in cui si nota come la FED abbia inizialmente incrementato i tassi, probabilmente assecondando la caduta dei prezzi, per poi procedere ad uno stimolo monetario abbassandoli dal 3% al 2,5%. Ciò ha ovviamente favorito il recupero. Il fatto tuttavia che tale stimolo si sia arrestato a inizio ottobre 2001 induce il modello a prevedere prezzi nuovamente in discesa, mentre nella realtà ciò non è avvenuto.

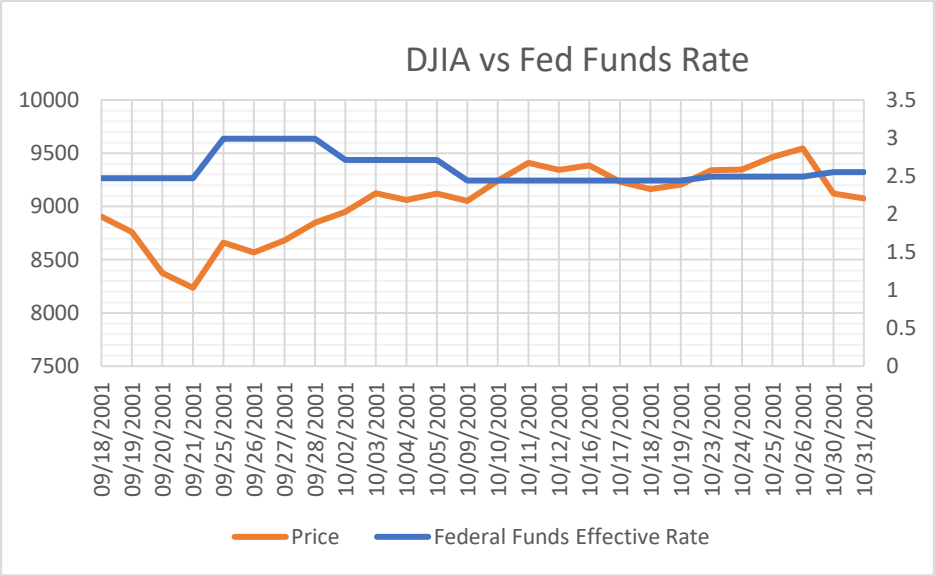


Figura 5: Confronto tra andamento predetto e andamento reale DJIA – Attacco alle Torri Gemelle (11 settembre)

Nel terzo scenario, incentrato sulla crisi dei mutui *subprime* (c.d. "crisi *Lehmann Brothers*" – Figura 6) il modello si accorda abbastanza bene con l'andamento reale nel tratto discendente della curva, andando tuttavia successivamente a sovrastimare la ripresa. Ciò probabilmente è dovuto al fatto che i poderosi stimoli monetari esercitati dalla FED per sostenere l'economia non hanno sortito un proporzionale effetto nella realtà, a differenza di quanto previsto dal modello, che come sopra descritto invece si aspetta evoluzioni proporzionali della variabile indipendente a fronte di azioni sulle variabili indipendenti (tra cui, appunto, i tassi di interesse).

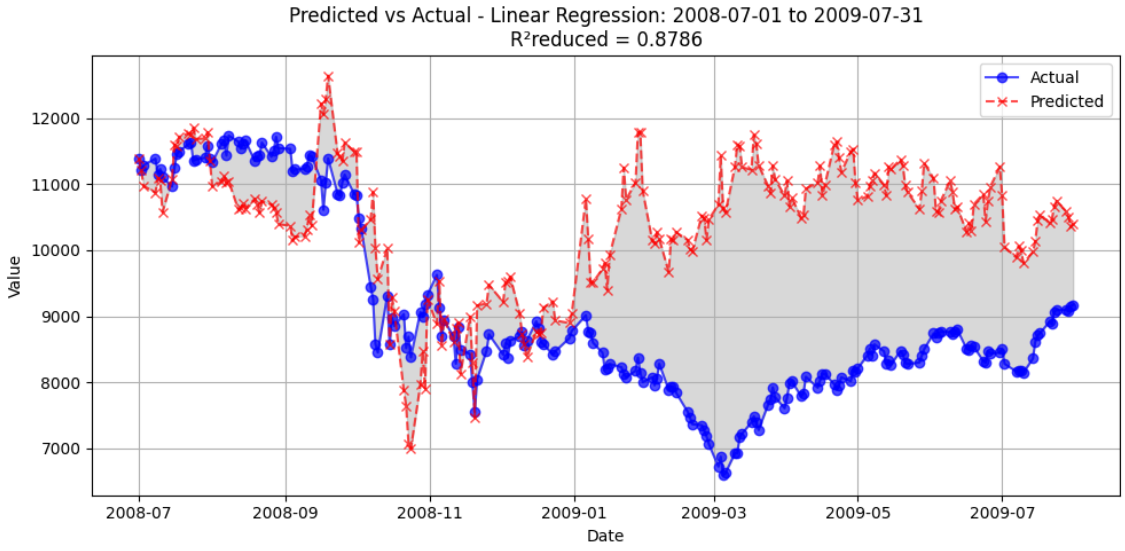


Figura 6: Confronto tra andamento predetto e andamento reale DJIA – Crisi Lehmann Brothers

In effetti l'applicazione delle politiche di stimolo della domanda aggregata per sostenere la ripresa dalle crisi finanziarie è una classica strategia di stampo *keynesiano*[10], fin dai tempi della Grande Recessione[11]. La logica

[10] Si vedano nuovamente i principi generali espressi da Sir John Maynard Keynes, "*The General Theory of Employment, Interest and Money.*",

[11] La grande crisi finanziaria del 1929, dalla quale gli USA si ripresero gradualmente adottando politiche di stimolo della domanda aggregata mediante imponenti programmi infrastrutturali e di spesa pubblica (il cosiddetto "*New Deal*" del presidente F.D. Roosvelt)

di tali interventi si fonda sull'assunto che, in presenza di una carenza di domanda aggregata, lo Stato possa svolgere un ruolo attivo nell'invertire il ciclo economico negativo. Questi interventi si realizzano principalmente attraverso due canali: l'aumento della spesa pubblica, mediante programmi di lavori pubblici, investimenti infrastrutturali e appalti statali, e l'espansione della base monetaria, attuata attraverso politiche monetarie accomodanti volte a sostenere i livelli dei prezzi, prevenire la deflazione e mitigare il rischio di insolvenze diffuse nel sistema economico[12]. Più volte, tuttavia, in tempi più recenti, tali strategie, si sono rivelate meno efficaci[13].

Il modello adottato, che registra il funzionamento implicito di tali politiche espansive nella generalità delle situazioni, in questi casi specifici prevede quindi un rialzo degli indici più sostenuto di quanto effettivamente riscontrato.

Nel caso di specie, le politiche monetarie espansive[14] della FED sono iniziate a novembre 2008 e sono proseguite per molto tempo, esaurendosi del tutto solamente nel 2015. Come si vede in Figura 6, tuttavia, immediatamente a valle della crisi, e per tutto il primo semestre del 2009, gli effetti sui prezzi azionari sono stati molto ridotti. Questo disallineamento suggerisce l'esistenza di fattori strutturali o contingenti che limitano l'efficacia delle politiche di stimolo, quali alti livelli di indebitamento preesistenti, un'elevata incertezza economica, il rischio di trappole della liquidità o l'erosione della fiducia dei consumatori e degli investitori, elementi che il modello non cattura se non in via indiretta.

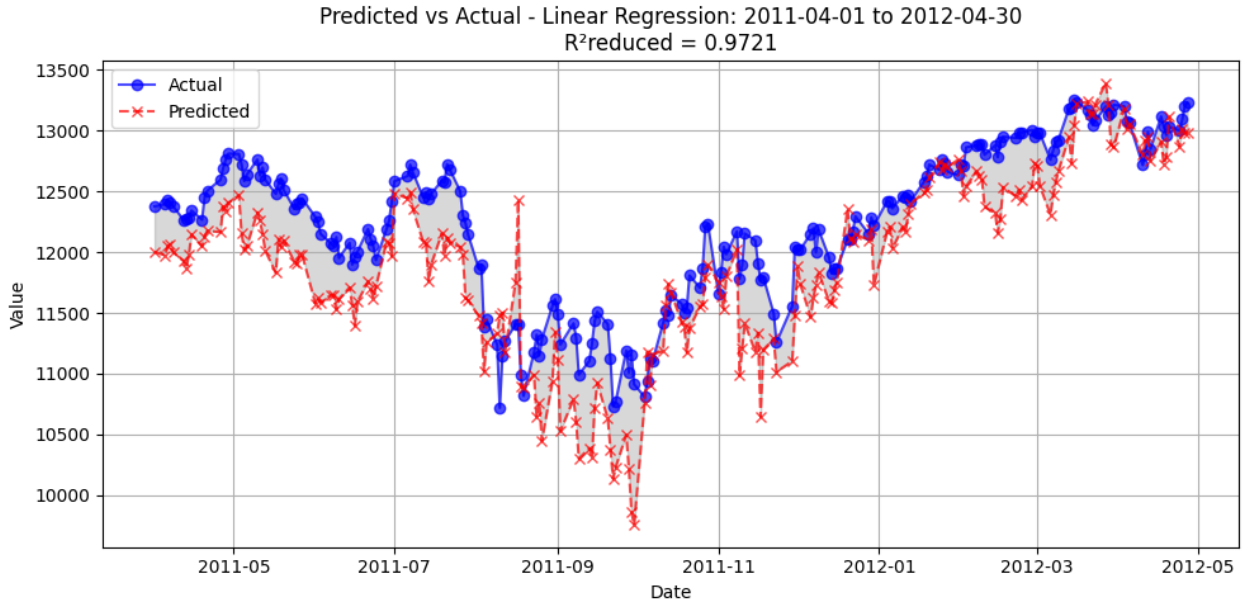


Figura 7: Confronto tra andamento predetto e andamento reale DJIA – Crisi del debito sovrano

Nel quarto scenario, ossia la "Crisi del debito sovrano", una delle dirette conseguenze delle politiche espansive di spesa pubblica e sostegno al debito praticate da molti Paesi impattati dalla crisi dei mutui *subprime*, il

[12] Un esempio estremo di tale tipo di stimolo monetario è la cosiddetta "*Helicopter Money*" teorizzata – provocatoriamente – da Milton Friedman in "*The Optimum Quantity of Money*" del 1969.

[13] Anche Keynes aveva previsto tale evenienza in certe condizioni particolari, la cosiddetta "trappola della liquidità", riferita al fenomeno per cui sotto un certo livello, l'abbassamento dei tassi non favorisce gli investimenti, ma bensì l'accumulo di liquidità.

[14] Il cosiddetto "*Quantitative Easing*", una formidabile espansione del passivo della FED derivante dall'acquisto di enormi quantità di titoli di debito ed obbligazioni sul mercato secondario per sostenere i prezzi e la solvibilità dei titoli. Si veda "*The New Tools of Monetary Policy*" di Ben S. Bernanke su *American Economic Review* 2020, 110(4): 943–983 e soprattutto il discorso tenuto alla *London School of Economics* da Ben S. Bernanke del 13/01/2009: "*The Crisis and the Policy Response*".

modello si accorda molto bene con l'andamento reale, pur manifestando un andamento meno "*smooth*" rispetto alla curva delle quotazioni effettive.

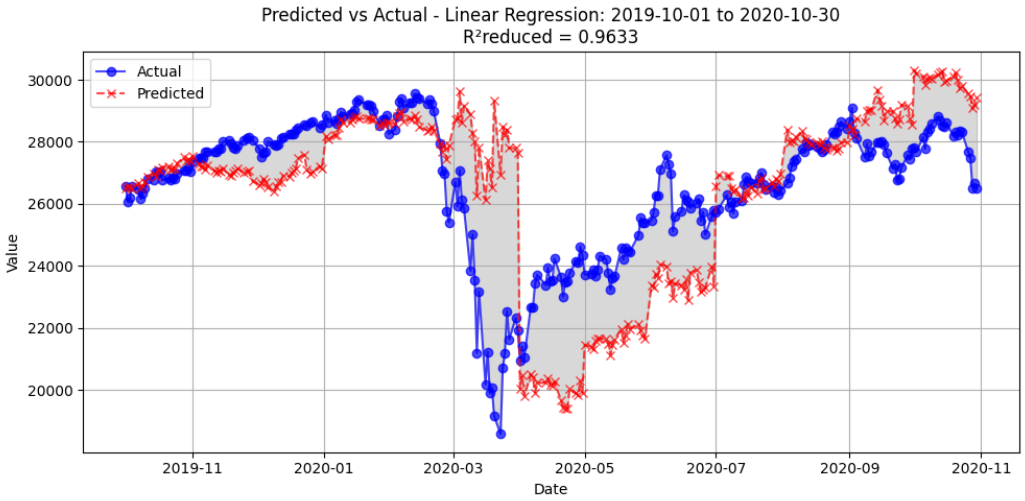


Figura 8: Confronto tra andamento predetto e andamento reale DJIA – Pandemia COVID-19

Nel quinto scenario, infine, il modello segue abbastanza bene la curva reale, pur mostrando una sorta di "ritardo", il che può essere spiegato tenendo conto che i mercati hanno reagito istantaneamente ad un evento del tutto alieno alle dinamiche finanziarie ed economiche, ossia le notizie legate al progressivo diffondersi della pandemia e le conseguenti azioni di restrizione delle attività umane nel tentativo di limitare il contagio.

Tutto ciò si è riverberato sui parametri economico-finanziari considerati dal modello in maniera non istantanea.

4. Correlazioni tra Caratteristiche

A partire dal modello completo, basato su 32 attributi, l'analisi delle relazioni tra le variabili indipendenti e la variabile *target* è stata condotta attraverso due strumenti principali: la *Correlation Line* (Figura 9) e la Matrice di Correlazione (Figura 10). Questi strumenti offrono una visione dettagliata delle interazioni tra le caratteristiche, evidenziando quelle più rilevanti per il modello predittivo.

4.1. Correlation Line

La *correlation line* (Figura 9) mostra il coefficiente di correlazione tra ogni variabile indipendente e la variabile *target* ("*Price*"), consentendo di identificare rapidamente quali variabili hanno una maggiore influenza predittiva.

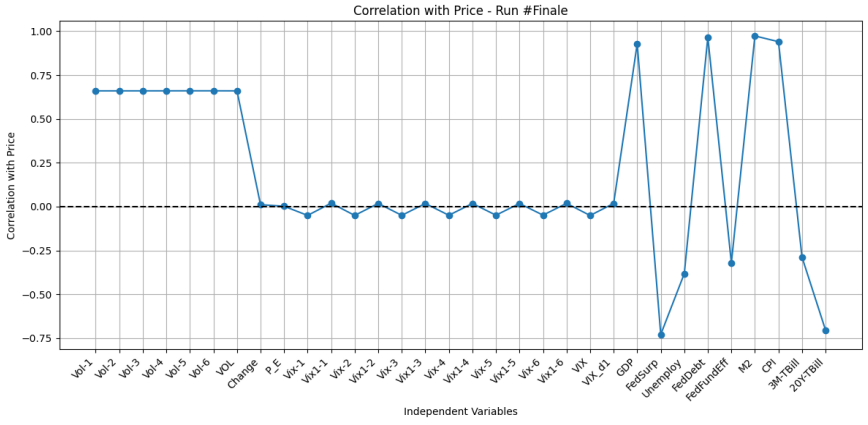


Figura 9: Correlation Line

I risultati principali sono i seguenti:

- **Volumi ritardati**: le variabili ritardate dei volumi ("Vol-1" fino a "Vol-6") presentano correlazioni elevate con la variabile *target*. Ciò conferma l'importanza delle dinamiche regressive lungo l'asse temporale.
- **VIX e Variazioni giornaliere VIX (Vix1)**: anche queste variabili mostrano una certa correlazione, sebbene inferiore rispetto ai volumi ritardati. Ciò suggerisce che la volatilità abbia un certo ruolo nella predizione della variabile *target*.
- **Indicatori macroeconomici**:
    - **Prodotto Interno Lordo (GDP)**: mostra una correlazione positiva significativa, suggerendo che una crescita economica robusta si riflette direttamente nei prezzi di mercato;
    - **Tasso base FED**: la correlazione negativa con la variabile *target* evidenzia l'effetto delle politiche monetarie sui prezzi;
    - **Debito Pubblico Totale**: presenta una correlazione positiva, indicando che alti livelli di indebitamento pubblico possono influire sui mercati[15].

4.2. Matrice di Correlazione

La matrice di correlazione (Figura 10) fornisce un'analisi più approfondita delle relazioni tra le variabili indipendenti stesse.

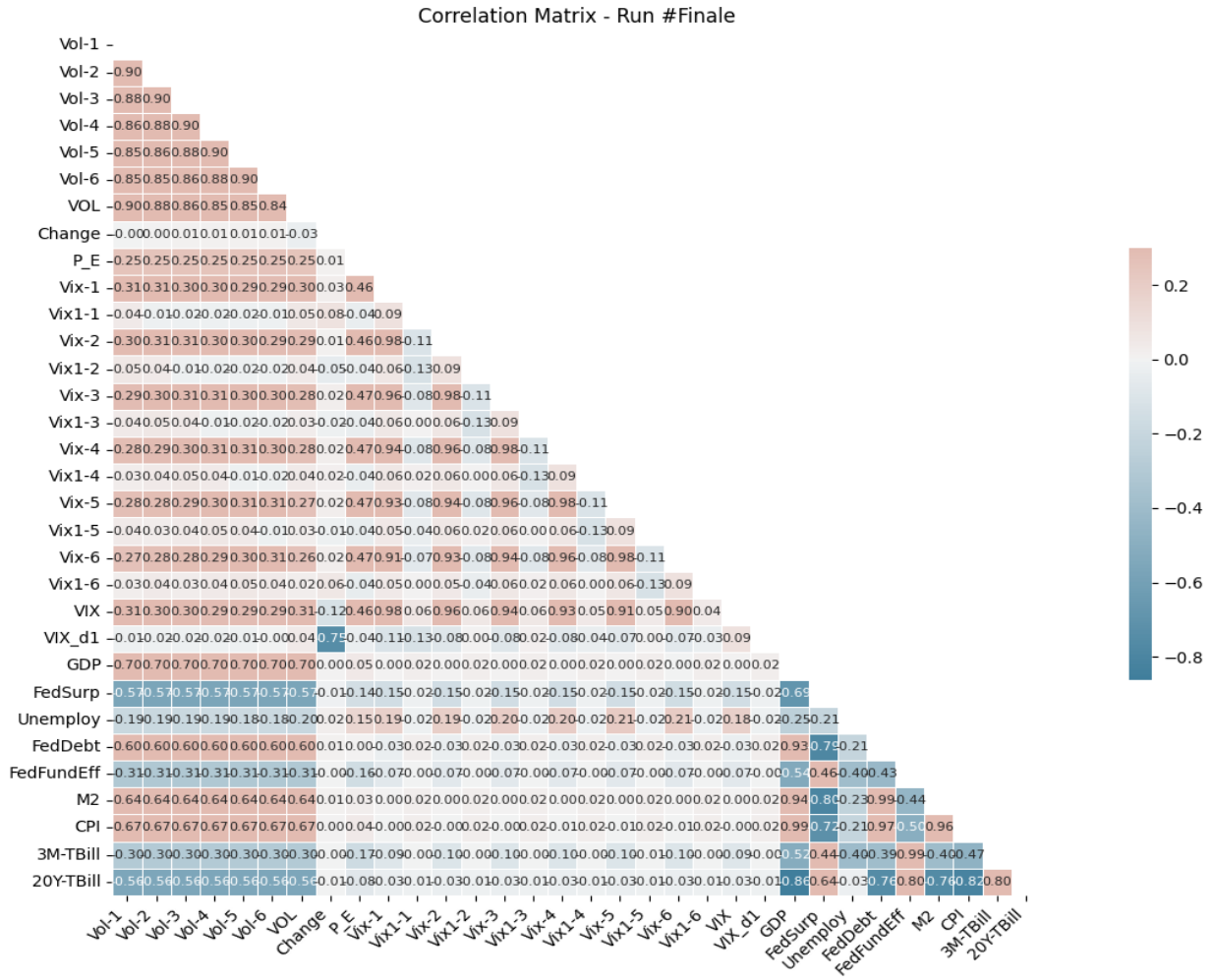


Figura 10: Matrice di Correlazione

[15] Si rammenti quanto analizzato al precedente paragrafo a proposito del ruolo di stimolo della spesa pubblica nei confronti della domanda aggregata attraverso programmi di lavori pubblici, investimenti infrastrutturali e appalti statali.

I principali spunti analitici sono:

- **Autocorrelazioni elevate tra parametri ritardati**: le variabili campionate su base giornaliera ripetuta mostrano correlazioni molto forti tra loro, confermando che i valori storici hanno un certa influenza sui valori futuri. Questo sottolinea l'importanza della componente temporale nella modellazione.
- **Indicatori di volatilità (VIX)**: le variabili relative alla volatilità (ad esempio, VIX e VIX_1st_derivative) mostrano correlazioni più deboli con le altre variabili e la variabile *target*. Tuttavia, il loro contributo potrebbe essere rilevante in condizioni di mercato altamente volatili.
- **Relazioni macroeconomiche trasversali**: variabili come GDP, M2 e CPI mostrano correlazioni significative tra loro, evidenziando la stretta connessione tra attività economica, liquidità e dinamiche inflazionistiche.

4.3. Osservazioni congiunte

L'analisi combinata di *correlation line* e matrice di correlazione fornisce una visione completa della struttura delle relazioni:

- le istanze delle variabili campionate lungo l'arco temporale presentano correlazioni elevate sia con la variabile *target* sia tra di loro;
- gli indicatori macroeconomici offrono un contributo cruciale per arricchire l'interpretabilità del modello, evidenziando le interazioni tra politica monetaria, liquidità e mercato;
- le variabili di volatilità, pur mostrando correlazioni più deboli, sono utili per migliorare la robustezza del modello in condizioni di mercato particolarmente instabili.

L'integrazione di queste analisi ha permesso di identificare le variabili più influenti per il modello predittivo, guidando la selezione delle caratteristiche.

5. Ottimizzazione del modello

Fra quelli considerati, il modello migliore si è rivelato finora quello basato sulla regressione lineare. Allo scopo di migliorare la sua qualità, bilanciando obiettivi concorrenti quali la semplicità di implementazione, la limitazione dell'estensione del *dataset*, la significatività delle variabili e la capacità del modello di separare il segnale dal rumore si è seguito un processo iterativo, con verifiche costanti sulle evidenze di multicollinearità, sul numero di variabili, sulla capacità di generalizzazione ed interpretazione del fenomeno, sulla robustezza agli errori di misurazione od ai valori mancanti, sulla significatività statistica e sulla coerenza con l'applicazione pratica.

Tradizionalmente, la costruzione del modello atto ad identificare la consistenza dei dati segue un approccio *ad hoc* che si basa su cicli ripetitivi di tentativi ed errori. Tuttavia, non vi è alcuna garanzia che il modello finale soddisfi tutti gli obiettivi desiderati in modo ottimale. Il nostro approccio mira ad eliminare la necessità di ripetere manualmente questi passi, fornendo un algoritmo che costruisca un set di modelli di alta qualità. Questo approccio bilancia gli obiettivi del modellatore, tra cui interpretabilità, parsimonia, robustezza ai dati rumorosi e significatività statistica, utilizzando l'ottimizzazione basata su CPLEX.

Quando il numero di possibili caratteristiche legato ai dati è elevato, spesso desideriamo identificare un sottoinsieme critico che sia principalmente responsabile della generazione della risposta. Ciò porta a modelli più interpretabili e migliora la precisione delle previsioni eliminando le variabili di rumore, aumentando così la capacità del modello di generalizzare. Per questa ragione, è spesso utile sviluppare modelli di regressione lineare con un numero massimo specificato k di coefficienti β diversi da zero. Questo numero k è chiamato "sparsità" del modello e rappresenta anche il numero massimo di predittori all'interno del modello.

Il modello utilizzato prevede di determinare il vettore β minimizzando il *Mean Square Error* (MSE) così definito:

$$\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}_\mathbf{o}||^2$$

dove:

- $\mathbf{X}$: è la matrice delle osservazioni;
- $\mathbf{y}$: è la variabile che si vuole predire.

Una relazione quasi-lineare tra le variabili indipendenti oscura la relazione di ciascuna caratteristica con la risposta del modello e porta a stime di parametri instabili. Per evitare questi problemi e produrre modelli interpretabili, un modello di regressione di alta qualità conterrà caratteristiche che sono il più ortogonali possibile. Pertanto, si è utilizzata la correlazione a coppie come misura di multicollinearità e costruito una sparsità selettiva limitando le variabili indipendenti nel modello di regressione a quelle che hanno una correlazione a coppie relativamente bassa. Questa è una tecnica standard che prevede di impiegare le variabili indipendenti con un valore di correlazione a coppie massimo tale per cui $C > C_{max}$=0,70[16] non dovrebbero essere incluse nell'analisi di regressione multipla. Altri metodi per gestire la multicollinearità includono la regressione a componenti principali e i minimi quadrati parziali[17] che trasformano i dati per produrre nuove variabili indipendenti non correlate. Sebbene questi risolvano efficacemente il problema della multicollinearità, potrebbe essere difficile interpretare le nuove caratteristiche e quindi non è chiaro fino a che punto le variabili originali influenzino la risposta. La regressione penalizzata, che fornisce stime distorte ma riduce la varianza, è un altro metodo comune per affrontare le varianze elevate derivanti dalla multicollinearità. Sebbene questo possa ridurre le varianze, la contrazione indotta da questi metodi non rende effettivamente i dati meno correlati, e quindi non lo consideriamo uno strumento appropriato per incoraggiare modelli interpretabili. In tale contesto, sulla base della matrice di autocorrelazione C tra tutte le variabili, e definito il valore di autocorrelazione massimo C_{max} per il modello sono state stralciate tutte le variabili i cui valori di correlazione coppie $C_{i,j} > C_{max} \forall i, j = 1..m$ definendo una nuova matrice di correlazione a coppie costituita da tutte le variabili che rispettano la citata condizione $\mathbf{HC}$.

Uno degli obbiettivi principali della ricerca è costruire modelli semplici ed interpretabili, specialmente considerando l'elevato numero di predittori in gioco prima dell'analisi dei dati. È risultato quindi particolarmente utile imporre un vincolo di sparsità, ossia un vincolo che consente di selezionare un sottoinsieme di variabili rilevanti per il modello, eliminando o riducendo il contributo di quelle meno significative. Questo approccio non solo migliora l'interpretabilità del modello, ma contribuisce anche a mitigare il rischio di sovradattamento (*overfitting*).

L'approccio utilizzato nell'elaborato per indurre la sparsità consiste nell'imporre un vincolo esplicito sul numero massimo di variabili indipendenti da includere nel modello. Tale vincolo detto **vincolo di sparsità basato su cardinalità** è definito come:

$$\sum_{i=1}^{m} z_i < k$$

dove:

[16] Si veda al proposito Tabachnick BG, Fidell LS (2001) *Using Multivariate Statistics 4th ed.* (Allyn and Bacon, Boston).
[17] Si veda al proposito Massy WF (1965) *Principal components regression in exploratory statistical research.* J. Amer. Statist. Assoc. 60(309):234–256.

- k è numero massimo di variabili indipendenti che si vogliono includere nel modello;
- z_i sono variabili booleane legate ai coefficienti di regressione β attraverso la relazione:

$$\begin{cases} \beta_i \leq z_i * M \\ \beta_i \geq -z_i * M \end{cases} \forall i = 1..m$$

con M molto grande tale per cui $\beta_i \in [-M, +M]$

L'interpretazione che ne discende è:

$$\begin{cases} z_i = 0 \; se \; \beta_i = 0 \\ z_i = 1 \; se \; \beta_i \neq 0 \end{cases}$$

Questo approccio è particolarmente utile in contesti dove è richiesta una selezione rigida delle variabili per motivi di interpretabilità o per limitare la complessità computazionale.

Definito il vettore $\mathbf{z}$ e la matrice di correlazione a coppie $\mathbf{HC}$ il vincolo introdotto per eliminare multicollinearita a coppie afferma che per ogni coppia di indici (i,j) presenti in HC,non è possibile selezionare contemporaneamente entrambi i predittori *i* e *j* ossia:

$$\forall < i, j > in \; \mathbf{HC} \; z_i + z_j \leq 1$$

quindi almeno uno dei due valori z_i o z_j deve essere uguale a 0.

5.1. Esiti finali dell'ottimizzazione

Per lo sviluppo del modello di ottimizzazione è stato preso in considerazione il *dataset* contenente tutte le osservazioni delle variabili indipendenti elencate nel capitolo 4. I valori delle osservazioni sono stati normalizzati in modo tale da riportare tutti i dati all'interno dell'intervallo [0,1].

Il vettore dei parametri β è stata individuato considerando come funzione obbiettivo la minimizzazione del MSE ed impiegando il 50% delle osservazioni (nTrain) random all'interno del *dataset* secondo la relazione:

$$\min_{\beta_i} \sum_{i=1}^{nTrain} (y_i - \sum_{j=1}^{m} \beta_j \, X_{i,j} - \beta_0)^2$$

con y_i variabile dipendente (precedentemente identificata come *"Price")* di cui si vogliono stimare i coefficienti *β* necessari per la predizione, mentre per verificare la bontà del modello di ottimizzazione è stato calcolato il **coefficiente di determinazione** R^2 utilizzando il 100% delle osservazioni (*Test*) assumendo che:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} \beta_j \, X_{i,j} - \beta_0)^2}{\sum_{i=1}^{n}(y_i - y_{bar})^2}$$

dove $y_{bar} = \frac{\sum_{i=1}^{n} y_i}{n}$ rappresenta il valore medio delle y_i.

I valori assunti per il modello sono i seguenti:

- $M = 100$;
- $m = 32$;
- $n = 7669$ (Numero delle Osservazioni);
- $y = Price$ (Predizione della variabile "*Price*");

Il modello è stato implementato 3 volte assegnando i seguenti valori massimi per il coefficiente di correlazione a coppie:

- $C_{max} = 0{,}7$;
- $C_{max} = 0{,}5$;
- $C_{max} = 0{,}3$.

impostando come parametro per modellare la sparsità il numero massimo di variabili indipendenti che si vogliono includere nel modello, ossia $k = m$ (il numero complessivo delle variabili indipendenti disponibili) per generalizzare quanto più possibile. Si prenderanno in considerazione solo le variabili indipendenti i cui valori di correlazione a coppie soddisfino la relazione $C_{i,j} > C_{max} \forall i, j = 1..m$.

I risultati conseguiti sono esposti di seguito.

Caso C_{max}=0,7

- Numero di variabili indipendenti incluse nel modello: $k = 15$.
- R^2= 0,977;

Risultati ottenuti:

Variabili Indipendenti (X)	β	z
Vol-6	0,205976	1
Change %	0,034902	1
P_E	0,016407	1
Vix1-1	-0,07231	1
Vix1-2	-0,07144	1
Vix1-3	-0,02909	1
Vix1-4	-0,08082	1
Vix1-5	-0,08229	1
Vix-6	-0,09623	1
Vix1-6	-0,01608	1
CBOE Volatility Index: VIX_1st_derivative	-0,04635	1
Federal Surplus or Deficit [-]	-0,15004	1
Unemployment Rate	-0,38545	1
Federal Debt: Total Public Debt	0,645177	1
Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis	-0,00731	1
β_0 *(Intercetta)*	*0,412343*	

Caso C_{max}=0,5

- Numero di variabili indipendenti incluse nel modello: $k = 14$;
- R^2= 0,975;

Risultati ottenuti:

Variabili Indipendenti (X)	$\boldsymbol{\beta}$	$\boldsymbol{z}$
Change %	0,034902	1
P_E	0,016407	1
Vix1-1	-0,07231	1
Vix1-2	-0,07144	1
Vix1-3	-0,02909	1
Vix1-4	-0,08082	1
Vix1-5	-0,08229	1
Vix-6	-0,09623	1
Vix1-6	-0,01608	1
CBOE Volatility Index: VIX_1st_derivative	-0,04635	1
Federal Surplus or Deficit [-]	-0,15004	1
Unemployment Rate	-0,38545	1
Federal Debt: Total Public Debt	0,645177	1
Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis	-0,00731	1
β_0 *(Intercetta)*	0,34	

Caso C_{max}=0,3

- Numero di variabili indipendenti incluse nel modello: $k = 12$;
- R^2= 0,974;

Risultati ottenuti:

Variabili Indipendenti (X)	$\boldsymbol{\beta}$	$\boldsymbol{z}$
Change %	0,034902	1
P_E	0,016407	1
Vix1-1	-0,07231	1
Vix1-2	-0,07144	1
Vix1-3	-0,02909	1
Vix1-4	-0,08082	1
Vix1-5	-0,08229	1
Vix-6	-0,09623	1
Vix1-6	-0,01608	1
CBOE Volatility Index: VIX_1st_derivative	-0,04635	1
Federal Surplus or Deficit [-]	-0,15004	1
Unemployment Rate	-0,38545	1
β_0 *(Intercetta)*	0,37	

Forzando al massimo il vincolo di sparsità, ponendo $k = 1$ si ottiene un modello davvero minimo, che prevede un unico attributo (M2) e consegue comunque un $R^2 = 0{,}946$, a ulteriore conferma della fortissima correlazione tra massa monetaria e prezzi in generale, e, quindi, nel caso particolare di analisi, prezzi dei titoli del DJIA. Gli andamenti relativi ai 4 casi analizzati sono illustrati sinotticamente nella seguente Figura 11.

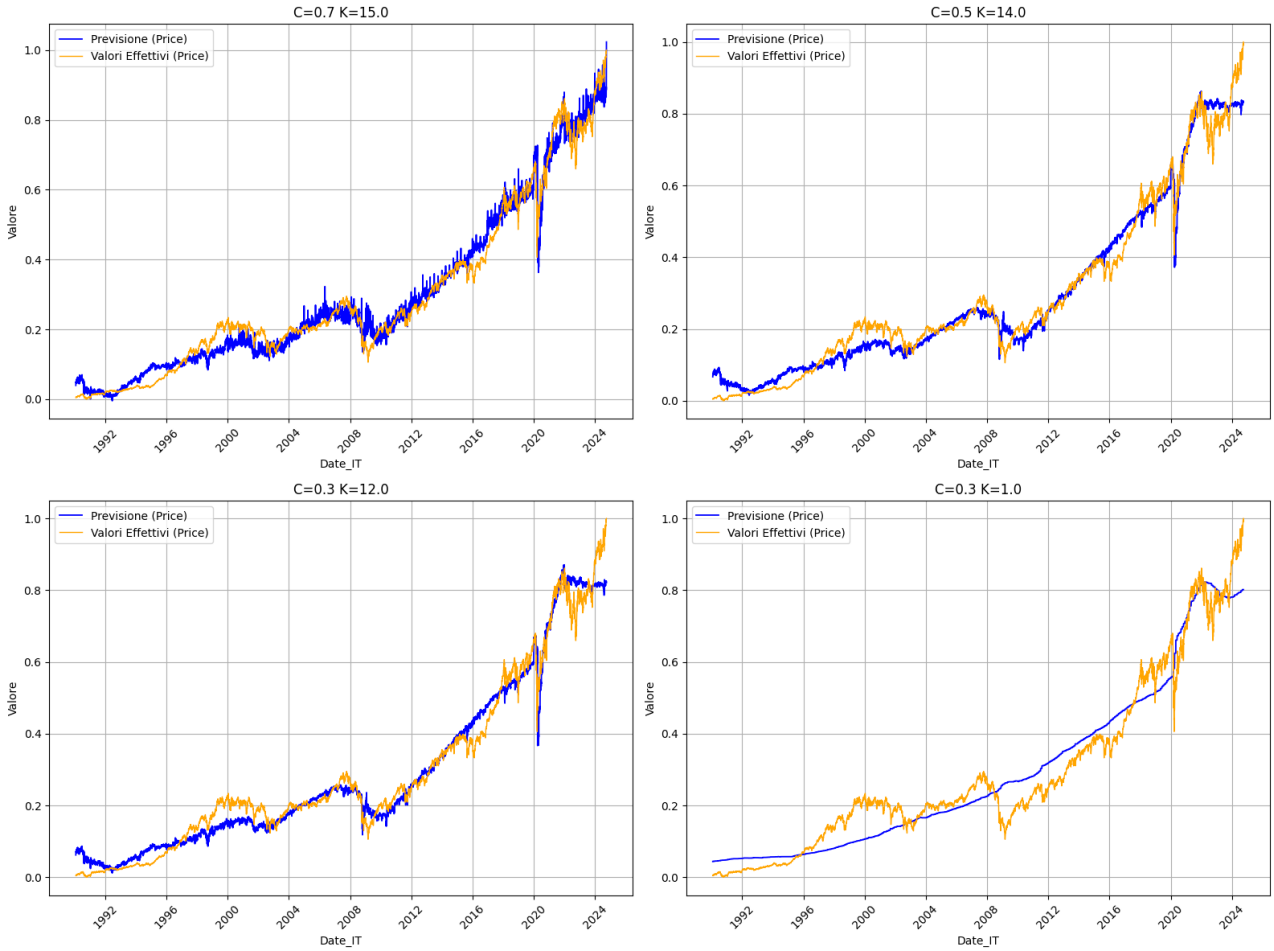


Figura 11: Confronto tra gli andamenti dei modelli a 15, 14, 12 e 1 variabile.

5.2. Ottimizzazione mediante massimizzazione del R^2 *Adjusted*

Il metodo sopra esposto, lascia una certa discrezionalità al decisore sulla base della scelta dei valori di $\boldsymbol{C_{max}}$ e $\boldsymbol{k}$. È possibile immaginare un approccio che automatizzi tale scelta. A tal fine, rimuovendo i vincoli di multicollinearità e sparsità, si è modificata la funzione obbiettivo in modo che i valori di $\boldsymbol{k}$ e del vettore dei coefficienti $\boldsymbol{\beta}$ vengano calcolati in modo da massimizzare la funzione $\boldsymbol{R^2_{Adjusted}}$. Tale funzione è una versione modificata del coefficiente di determinazione $\boldsymbol{R^2}$, definita per tenere conto del numero di variabili indipendenti (o predittori) nel modello di regressione. Generalmente, si utilizza per valutare la bontà del *fit* del modello tenendo conto della sua complessità (ovvero, del numero di predittori). Esso è particolarmente utile per la *feature selection* ossia l'identificazione delle variabili indipendenti più rilevanti per il modello, e per modelli predittivi complessi in cui si vuole evitare di includere troppe variabili che potrebbero portare a fenomeni di *overfitting* o modelli eccessivamente complicati. La formula per calcolare $\boldsymbol{R^2_{Adjusted}}$ è la seguente:

$$R^2_{Adjusted} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Mentre R^2 misura la frazione della devianza del *target* spiegata dal modello e tende ad aumentare con l'aggiunta di nuove variabili indipendenti, anche se non contribuiscono realmente a migliorare il modello, $R^2_{Adjusted}$ introduce una penalizzazione per l'aggiunta di nuove variabili indipendenti e non aumenta

necessariamente con più predittori; cresce solo se le nuove variabili migliorano significativamente il modello e può diminuire se si aggiungono variabili che non apportano valore predittivo. Infatti:

- Se si aggiungono variabili irrilevanti, il denominatore aumenta, causando una diminuzione di $R^2_{Adjusted}$;
- Se si aggiungono variabili rilevanti, il miglioramento di R^2 sarà sufficiente a compensare l'aumento di k e $R^2_{Adjusted}$ crescerà.

$R^2_{Adjusted}$ è quindi una funzione che tende a crescere con l'aggiunta di variabili rilevanti fino a quando l'aggiunta di ulteriori predittori non migliora R^2 abbastanza da compensare la penalizzazione dovuta all'aumento di k. In questa fase, $R^2_{Adjusted}$ potrebbe stagnare o diminuire. Per questi motivi la funzione $R^2_{Adjusted}$ è una funzione concava che avrà un punto di massimo in prossimità della quale il numero di predittori *k* sarà ottimo. È quindi possibile impostare un modello di ottimizzazione avente tale funzione come funzione obiettivo da massimizzare.

L'interpretazione del valore assunto da $R^2_{Adjusted}$ è di seguito esplicata:

- <u>Valori prossimi ad 1</u>
  la varianza del *target* viene quasi interamente spiegata e ha un numero di predittori ben bilanciato;
- <u>Valori prossimi a 0</u>
  non spiega tutta la varianza del *target* e ha troppi predittori rispetto al numero di dati disponibili.

Per l'implementazione del modello, è stato in prima analisi impiegato il compilatore CPLEX implementando la seguente funzione:

$$\max_{\mathbf{\beta},\mathbf{z}}\left(1-\frac{1-\frac{\sum_{i=1}^{nTrain}\left(y_i-\sum_{j=1}^{m}\beta_j X_{i,j}-\beta_0\right)\cdot(nTrain-1)^2}{\sum_{i=1}^{nTrain}\left(y_i-y_{bar}\right)^2}}{n-\sum_{j=1}^{m}z_j-1}\right.$$

Tuttavia per la presenza di $\sum_{j=1}^{m} z_j$ la funzione risulta non lineare, rendendo il problema non risolvibile tramite CPLEX.

Per superare tali difficoltà implementative, la ricerca dei massimi della funzione $R^2_{Adjusted}$ è avvenuta impiegando Python. La complessità computazionale del problema di ottimizzazione che si sta considerando è pari a:

$$Complex = 2^m - 1 = 4.294.967.295$$

Ciò aumenta enormemente i tempi di calcolo impendo di risolvere il modello in tempi accettabili. Si è quindi agito riducendo la complessità del sistema – e di conseguenza i tempi di calcolo – utilizzando **l'Algoritmo *Greedy Stepwise*** il quale, invece di calcolare tutte le possibili combinazioni dell'intero *dataset*, usa un approccio iterativo aggiungendo una variabile alla volta e scegliendo quella che migliora maggiormente $R^2_{Adjusted}$, interrompendosi quando l'inclusione di nuove variabili non migliora più significativamente R^2. Tale processo fissa l'ordine progressivo di aggiunta delle variabili, irrigidendone la composizione. Questa limitazione comporta il fatto che, laddove esistano combinazioni alternative, non inclusive delle variabili già selezionate ma magari caratterizzate da accuratezza migliore, esse non vengono individuate dall'algoritmo.

D'altro canto, il grosso vantaggio è che la complessità computazionale con l'algoritmo *Greedy Stepwise* è decisamente migliore del caso precedente e pari a:

$$Complex = m * n = 247.360$$

Con quest'ultimo approccio di ***feature selection*** è stato quindi determinato il seguente insieme di k=5 variabili:

Variabili Indipendenti (X)	$\boldsymbol{\beta}$
M2	0,5
Unemployment Rate	-0,363
Federal Surplus or Deficit [-]	-0,158
Consumer Price Index for All Urban Consumers: All Items in U.S. City Average	0,174
CBOE Volatility Index: VIX	-0,081
β_0 *(Intercetta)*	0,26

in corrispondenza delle quali si ha un valore di $R^2 = 0{,}976$, effettivamente migliore di quello precedentemente individuato quantomeno per il modello a 12 variabili, anche se comunque inferiore a quelli conseguiti dai modelli a più variabili.

Con l'elaborazione di questa nuova configurazione per gli attributi del modello di regressione lineare è possibile compilare la seguente tabella riepilogativa:

k (n. attributi)	**C_{max}**	**$R^2_{adjusted}$**	**R^2**	**Software utilizzato**	**Metodo di Ottimizzazione**
32	1	0,983004	0,982933	Python (*scikit-learn*/Regress. Lineare)	Min **MSE**
15	0,7	0,977045	0,977330	CPLEX	Min **MSE** con vincolo su **C_{max}**
12	0,3	0,974051	0,974570	CPLEX	Min **MSE** con vincolo su **C_{max}**
5	//	0,976016	0,976016	Python (*Greedy Stepwise*)	Max **$R^2_{adjusted}$**
1	//	0,946007	0,946007	CPLEX	Max $C(y/X_j)$[18]

I risultati rappresentati in tabella evidenziano come l'algoritmo *Greedy Stepwise* abbia individuato un possibile massimo locale per $k = 5$, i valori di $R^2_{Adjusted}$ calcolati puntualmente per i diversi valori di $k = 32, 15, 12, 5 \text{ ed } 1$, ci indicano che il modello di previsione ottimo – quantomeno in considerazione del $R^2_{adjusted}$ – è quello caratterizzato da 32 variabili.

[18] Massimizzazione della correlazione tra variabile dipendente e singola variabile indipendente.

5.3. Risultati post ottimizzazione

I modelli sono a questo punto stati modificati e riaddestrati conformemente alle indicazioni ottenute dall'ottimizzazione. Nella tabella seguente sono sintetizzati i risultati.

	Valore soglia corr.	R2	Max. Var. Ind.	Beta 0	Vol-1	Vol-2	Vol-3	Vol-4	Vol-5	Vol-6	Vol	Ch.%	P_E	Vix-1	Vix1-1	Vix-2	Vix1-2	Vix-3	Vix1-3	Vix-4	Vix1-4	Vix-5	Vix1-5	Vix-6	Vix1-6	VIX	VIX 1st derivative	GDP	Fed. Surplus or Deficit	Unempl. Rate	Fed. Debt: Total PubL. Debt	Fed Funds Effect. Rate	M2	CPI	Market Yield U.S. Treasury Sec. 3-Month	Market Yield U.S. Treasury Sec. 20-Year
CPLEX		0,97401	12		0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	0	0	0
	0,3			0,372962	0	0	0	0	0	0	0	0,039	0	0	-0,0721	0	-0,076	0	-0,0403	0	-0,0829	0	-0,0778	-0,0703	-0,0131	0	-0,0355	0	-0,1049	-0,3255	0	0	0,678531	0	0	0
Mod. Python		0,97401		0,385237								0,023			-0,0684		-0,0484		-0,0666		-0,07756		-0,0869	-0,0685	-0,011		-0,03402		-0,1083	-0,3328			0,675485			
CPLEX		0,97457	14		0	0	0	0	0	0	0	1	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	1	0	1	1	0	0	0
	0,5			0,339838	0	0	0	0	0	0	0	0,0377	0,028232	0	0,07143	0	-0,0725	0	-0,0403	0	-0,0817	0	-0,0769	-0,0858	-0,0081	0	-0,03694	0	-0,0909	-0,2978	0	0,020	0,697928	0	0	0
Mod. Python		0,97450		0,348565								0,0263	0,036257		-0,0117		-0,0998		-0,0518		-0,0773		-0,0626	-0,0759	-0,0452		-0,03901		-0,09507	-0,30025		0,024	0,693826			
CPLEX		0,97733	15		0	0	0	0	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	1	1	0	0	0	1	0
	0,7			0,412343	0	0	0	0	0	0,206	0	0,0349	0,016407	0	-0,0723	0	-0,0714	0	-0,0291	0	-0,0809	0	-0,0823	-0,0962	-0,0161	0	-0,04635	0	-0,15	-0,3855	0,645177	0	0	0	-0,0073078	0
Mod. Python		0,97737		0,440392						0,202		0,0407	0,005019		-0,0351		-0,078		-0,0794		-0,1056		-0,0772	-0,0785	-0,0421		-0,04446		-0,1572	-0,3834	0,644983				-0,00560376	
CPLEX																																				
	MOD. TOT.																																			
Mod. Python		0,982933178	32	0,102268	0,03884	0,00909	0,01219	0,02036	0,04786	0,047	0,023120	0,0191	0,035055	0,02094	-0,05791	-0,06546	-0,0250	-0,0035	-0,0227	0,04424	-0,027	-0,04792	-0,0596	-0,0146	-0,0142	-0,062	-0,01185	0,545	-0,0605	-0,2018	0,61736	0,273	0,207188429	-0,55859938	-0,2575717	0,128904108

Tabella 2: Confronto tra le varie configurazioni testate per i modelli di regressione lineare multivariata

Sia i β che i valori di R^2 calcolati tramite CPLEX o tramite i modelli implementati tramite *scikit-learn* sono piuttosto allineati.

I parametri fondamentali minimi che consentono di predire il valore del DJIA nel modello di regressione lineare multivariata implementata sono:

- Ch.%;
- Vix1-1;
- Vix1-2;
- Vix1-3;
- Vix1-4;
- Vix1-5;
- Vix-6;
- Vix1-6;
- VIX 1st derivative;
- Fed. Surplus or Deficit;
- Unempl. Rate;
- M2.

In estrema sintesi i parametri che spiegano meglio l'andamento dell'indice considerato sono:

- la M2, che rappresentando la liquidità monetaria del sistema economico di fatto costituisce il "carburante" che alimenta tutti i prezzi, e di fatti assume coefficiente con segno positivo;
- il tasso di disoccupazione, che costituisce il termometro dello stato di salute del sistema economico e dell'efficienza dello stesso[19], e infatti assume coefficiente con segno negativo;
- la variazione percentuale giornaliera dell'indice, che rappresenta in qualche modo la tendenza presente del mercato alla crescita o alla contrazione (coefficiente con segno positivo);
- l'intera serie autoregressiva delle derivate dell'indice VIX, con coefficiente positivo, che danno conto della volatilità del sistema e quindi della sua tendenza nella settimana antecedente a muoversi o a restare stabile.

Si nota immediatamente come il R^2 non vari molto al crescere delle variabili indipendenti utilizzate. Esso assume il valore di 0,982933178 con 32 variabili e 0,97401 con 12 variabili. Gli andamenti complessivi sono mostrati di seguito. L'ottimizzazione condotta ha quindi semplificato moltissimo il modello, senza sacrificarne eccessivamente la rappresentatività espressa in termini di R^2. Nella seguente Figura 12 il confronto tra gli andamenti generali predetti (confrontati con gli andamenti effettivi) in utilizzando i modelli rispettivamente a 32, 15 e 12 variabili.

[19] In Keynes, il tasso di disoccupazione rappresenta un indicatore centrale della sottoutilizzazione delle risorse produttive di un'economia, in particolare della forza lavoro, e riflette un disequilibrio nella domanda aggregata.

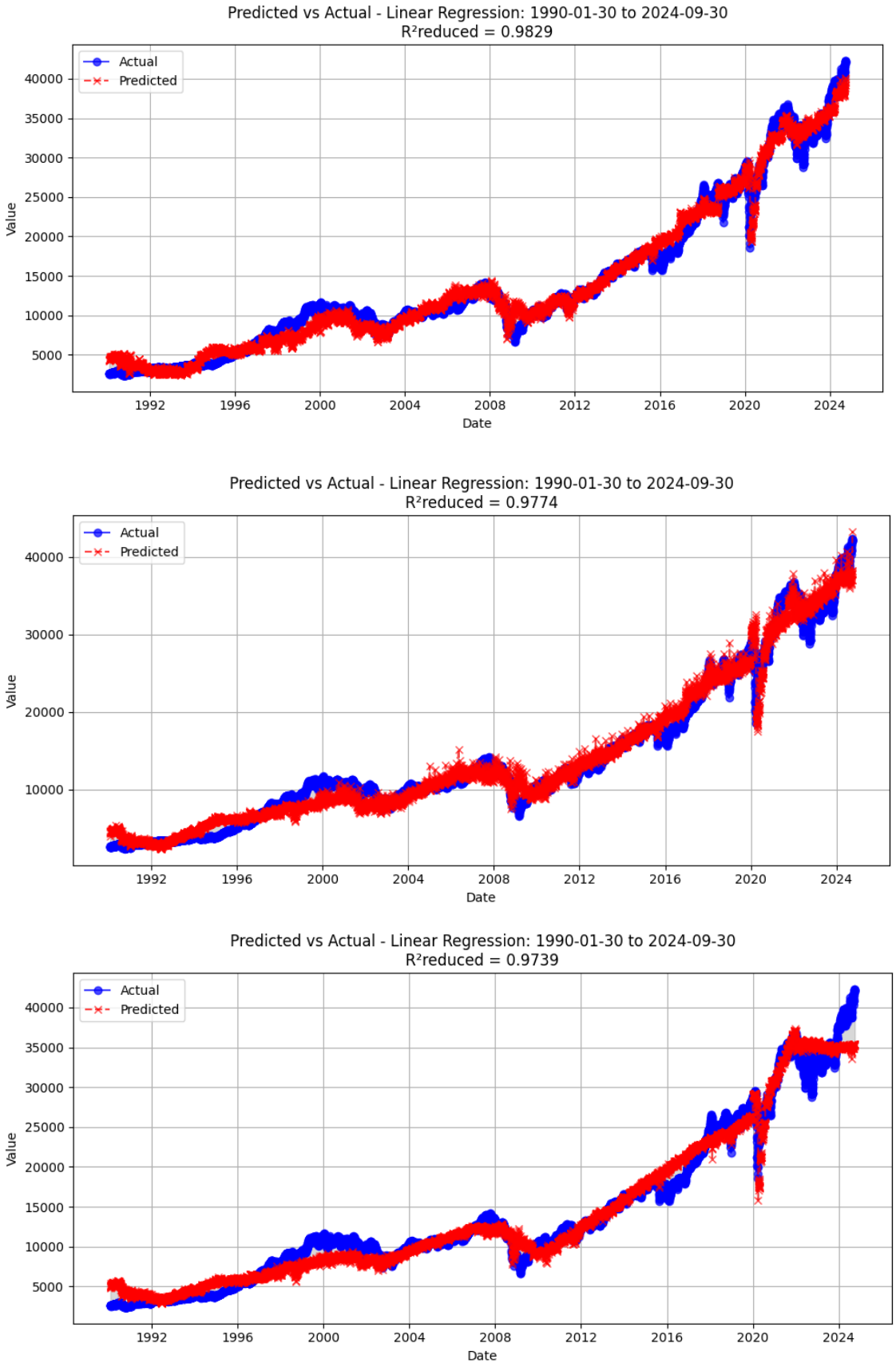


Figura 12: Confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.

Andando a paragonare gli andamenti previsti ed effettivi in occasione delle grandi crisi osserviamo la seguente situazione.

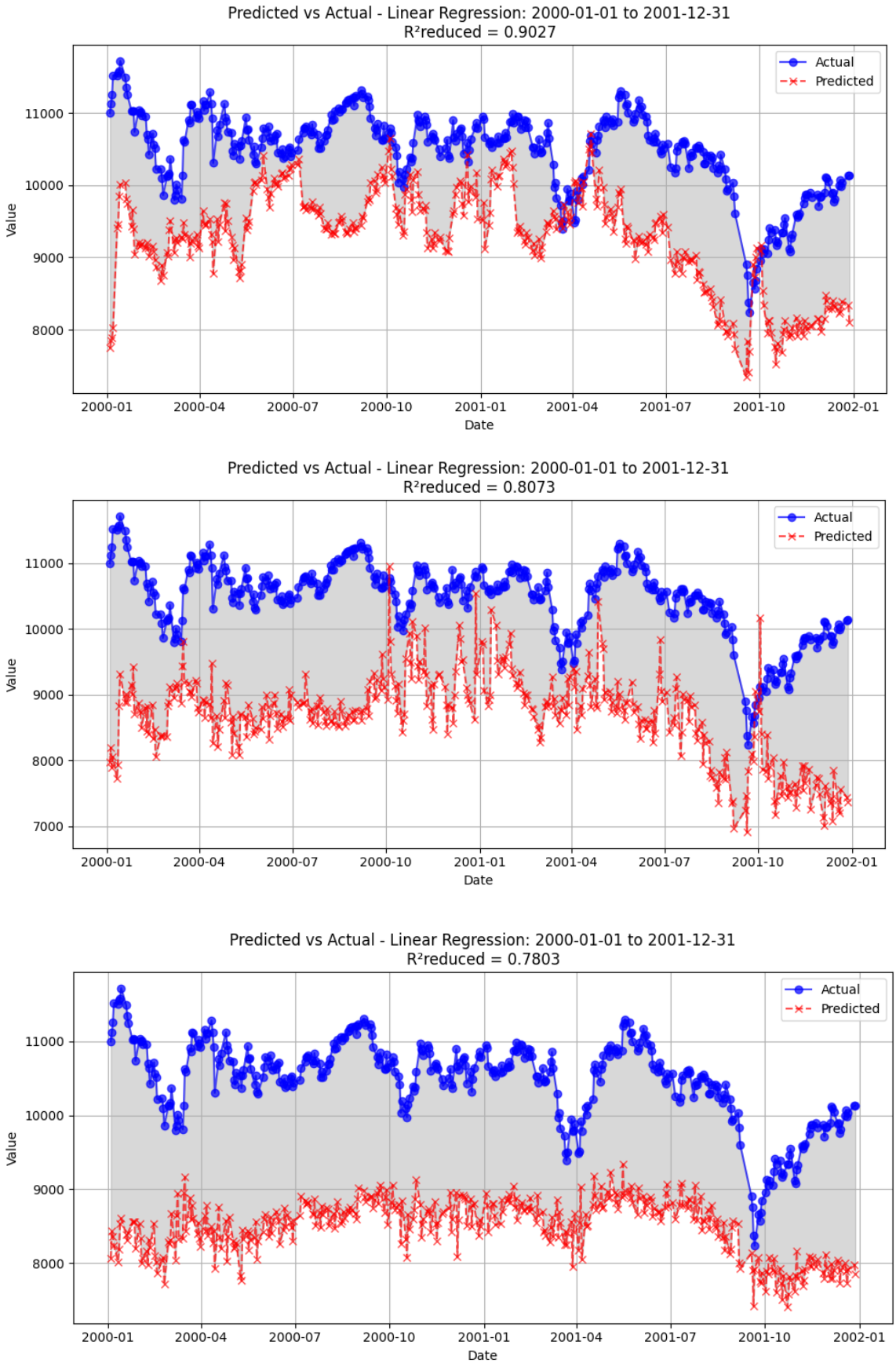


Figura 13: Crisi dot.com: confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.

Nello scenario "*dot.com*" la perdita di accuratezza tra il modello esteso e quelli a 15 e 12 variabili è evidente.

Non avviene la stessa cosa nello scenario "*11 settembre*" (l'accuratezza non varia molto tra i tre modelli); curiosamente il modello a 15 variabili è meno accurato di quello a 12 variabili.

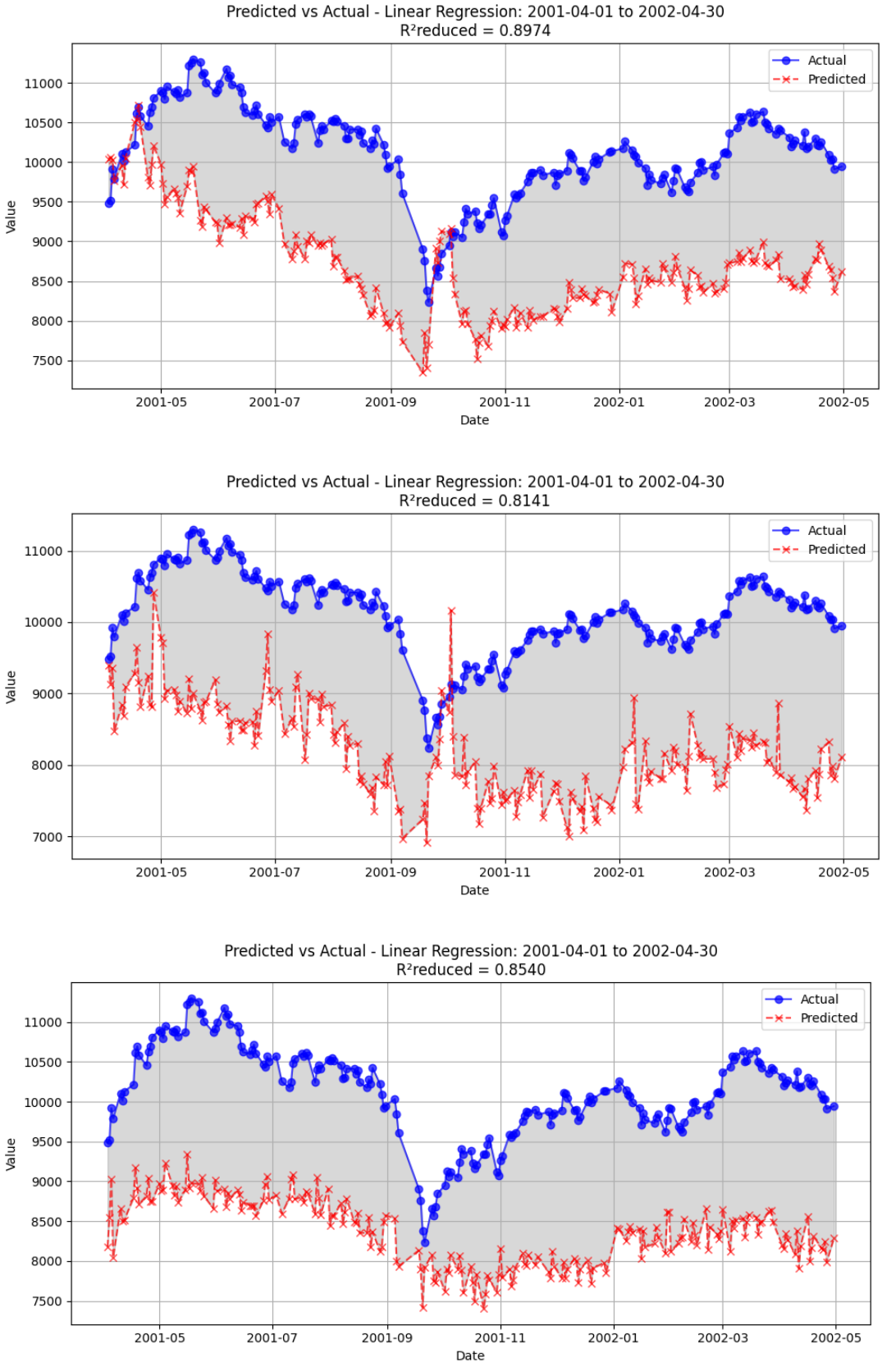
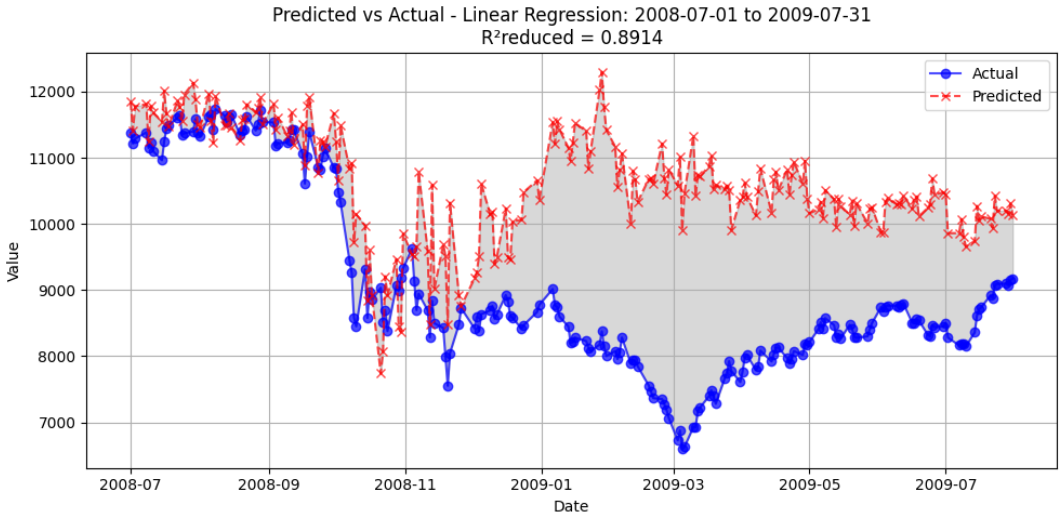
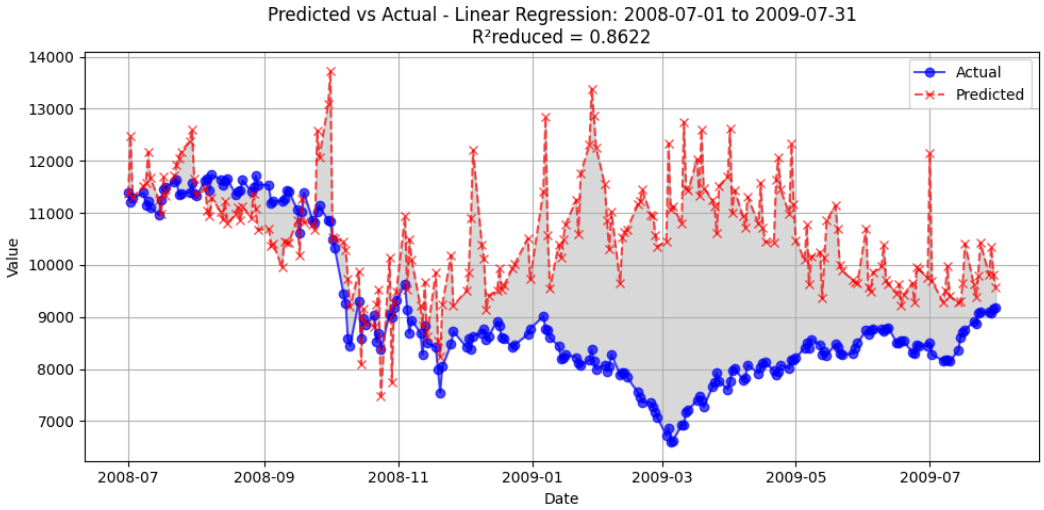


Figura 14: Attacco alle Torri Gemelle: confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.





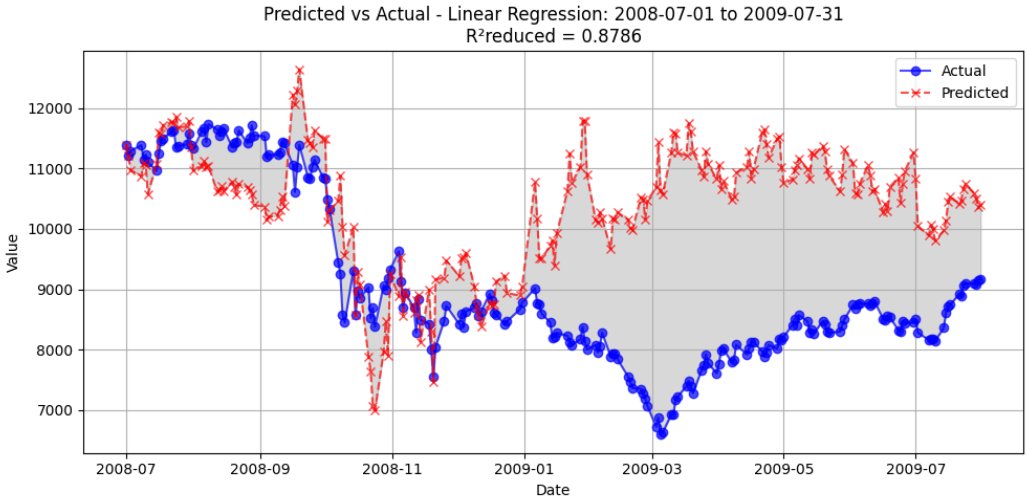


Figura 15: Lehmann Brothers: confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.

Nello scenario "*Lehmann Brothers*" nuovamente vi è una certa perdita di accuratezza tra il modello esteso e quelli a 15 e 12 variabili; la fase post crisi in particolare non viene rappresentata adeguatamente.

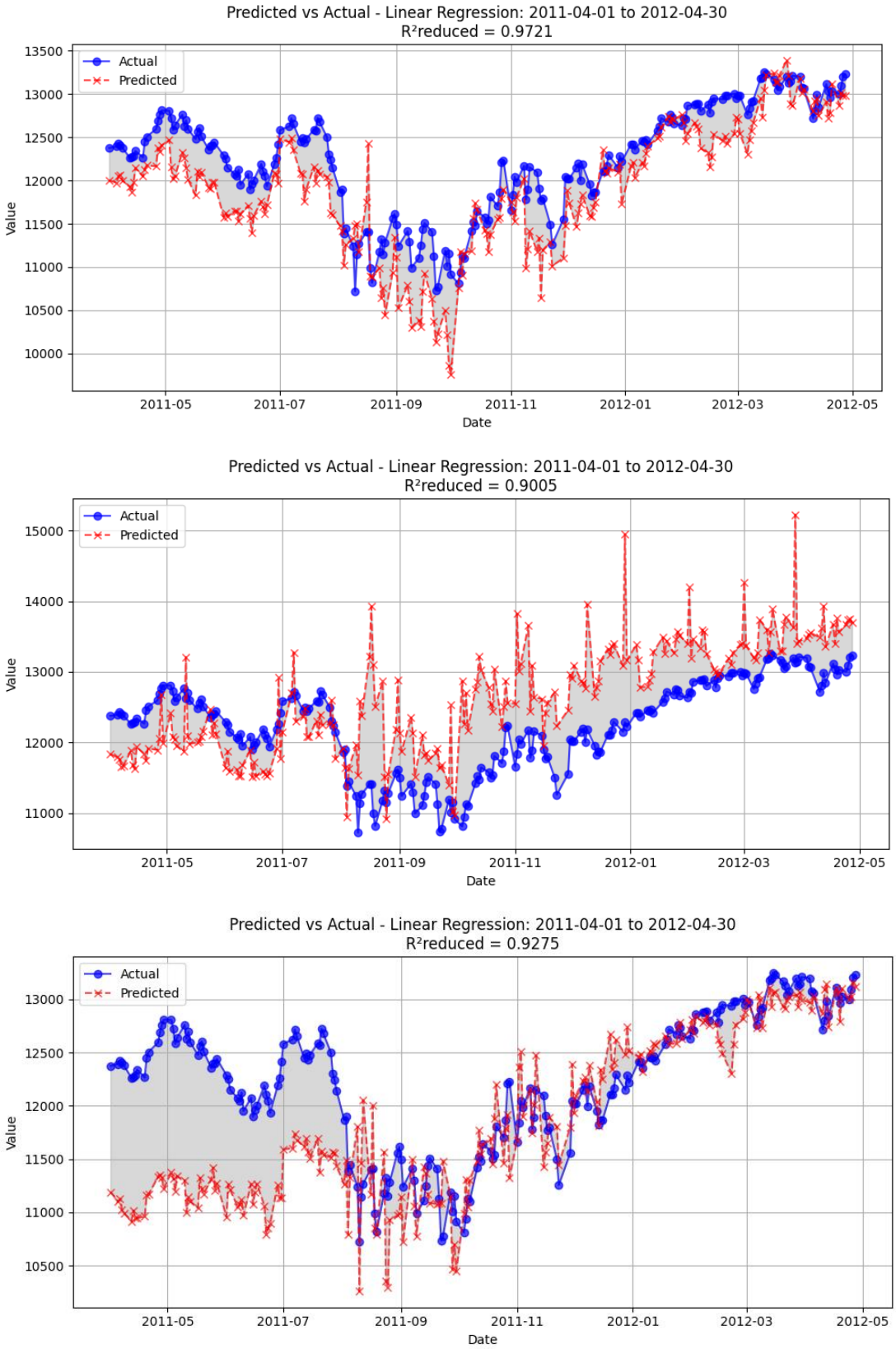
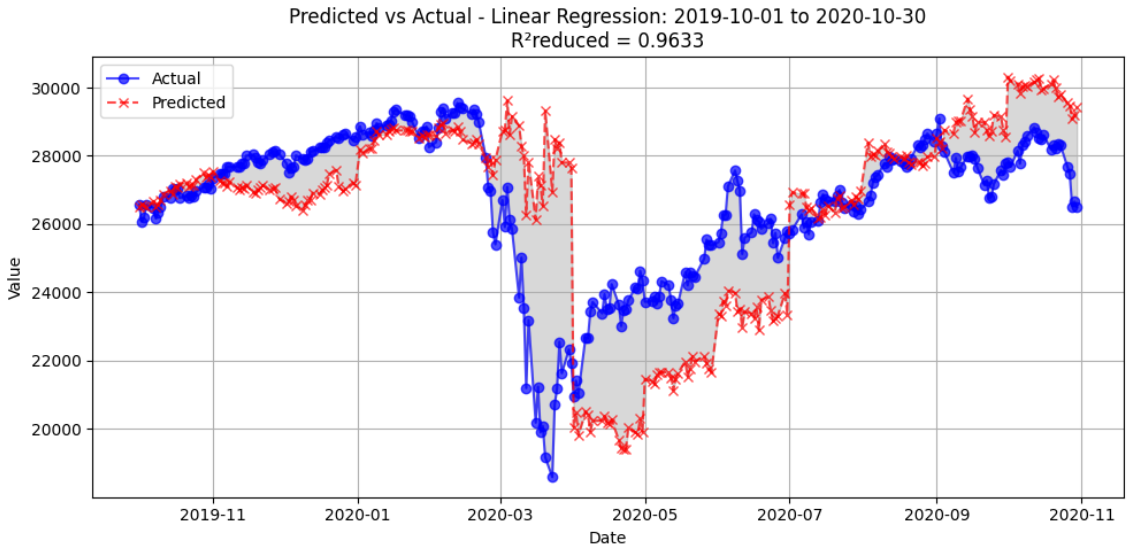
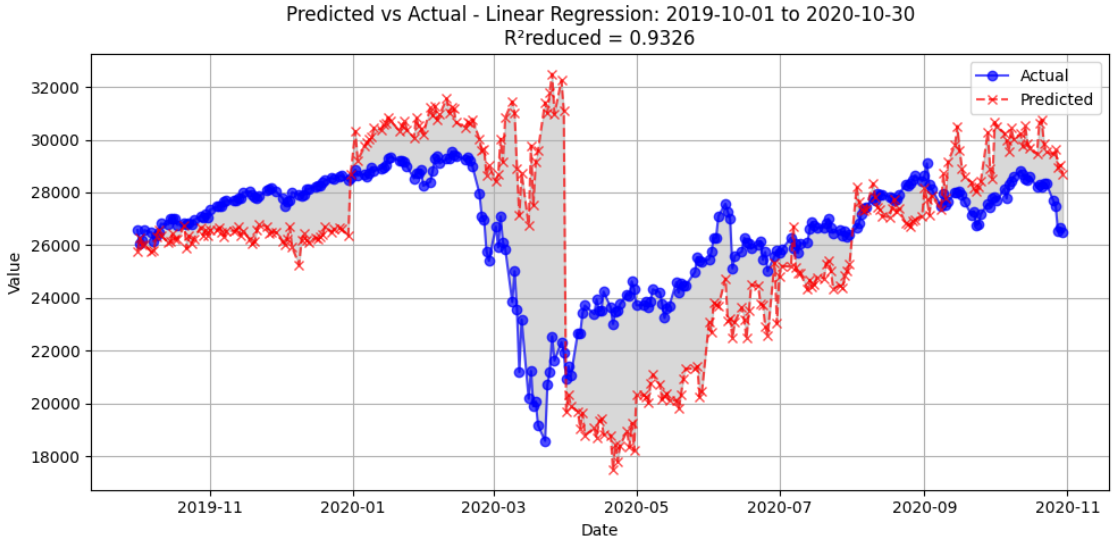


Figura 16: Crisi del debito sovrano: confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.

Nuovamente il modello a 15 variabili appare meno accurato di quello a 12 variabili.





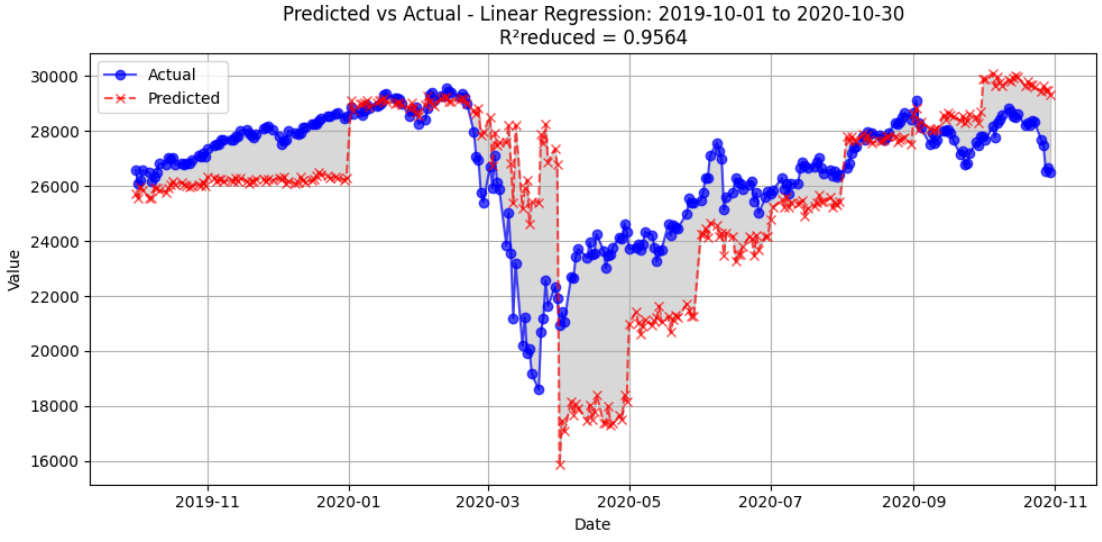


Figura 17: Pandemia COVID-19: confronto tra andamenti predetti e andamento reale complessivo – Modello a 32, a 15 ed a 12 variabili.

Anche in quest'ultimo caso il modello a 15 variabili appare meno accurato di quello a 12 variabili.

In ogni caso l'accuratezza del modello è peggiorata in tutti gli scenari, talora in modo molto consistente.

6. Interpretazione e predizione

Quanto finora mostrato consente di fare alcune considerazioni. Osserviamo esclusivamente le variabili incluse dopo l'ottimizzazione ed i relativi β, ossia:

Categoria	Parametro	β	INTERPRETAZIONE
VARIAZIONI	***Ch.%***	0,023000	ASPETTATIVE: CONTRIBUTI POSITIVI
VOLATILITA' DIFFERENZIALE	***Vix1-1***	-0,068390	ASPETTATIVE: CONTRIBUTI NEGATIVI
	Vix1-2	-0,048420	
	Vix1-3	-0,066610	
	Vix1-4	-0,077560	
	Vix1-5	-0,086860	
	Vix1-6	-0,011010	
	VIX 1st derivative	-0,034020	
VOLATILITA'	***Vix-6***	-0,068520	
PARAMETRI FINANZIARI	***Fed. Surplus/Deficit***	-0,108260	DINAMICA FONDAMENTALE
	M2	0,675485	
PARAMETRI MACROECONOMICI	***Unempl. Rate***	-0,332710	

Tabella 3: Parametri del modello minimo (in base all'ottimizzazione su MSE con vincolo su C_{max} - modello a 12 variabili) ed interpretazione

Possiamo ipotizzare che il modello di regressione lineare utilizzato consideri implicitamente una tendenza di fondo molto robusta dettata dai parametri macroeconomici e finanziari, su cui si innestano cicli di breve periodo dettati dalla combinazione degli effetti delle variazioni giornaliere (**Ch.%**) più alcuni valori assunti dalla **VIX** e soprattutto da tutte le sue variazioni giornaliere nei 7 giorni antecedenti.

Quanto sopra rispecchia le teorie economiche generali sulla formazione dei prezzi[20] menzionate nell'*Introduzione*, con le variabili responsabili dei cicli a breve che rappresentano le stime delle "aspettative"[21] degli investitori, che nel caso in esame si concentrano sulle possibilità che i prezzi delle azioni crescano, ed alle quali la **Ch.%** contribuisce positivamente come ipotesi proiettiva di un trend auspicabilmente crescente, mentre la **VIX** e le sue derivate vi contribuiscono negativamente, abbassando le aspettative in occasione di VIX crescenti e incrementandole laddove vi sia una tendenza di VIX decrescenti. Ciò è pienamente in linea con quello che la VIX effettivamente rappresenta e con le grandezze a partire dalle quali la VIX viene calcolata. Essa infatti è un indice in tempo reale che rappresenta le aspettative del mercato sulla forza relativa delle variazioni di prezzo a breve termine dell'indice S&P 500, e si ricava a partire dai prezzi delle opzioni sull'S&P 500 con scadenza a breve termine.

Tale interpretazione viene ulteriormente rinforzata considerando i modelli più ampi rispetto al minimo (quelli definiti impostando come soglia di correlazione almeno 0.5), che includono il parametro **P_E**, ossia il *Price/Earnings ratio* (rapporto Prezzo/Utile), che è una rappresentazione sostanziale delle aspettative degli investitori, poiché mostra il prezzo che si è disposti a pagare subito per partecipare alla distribuzione degli utili, e chiaramente è maggiore laddove ci si attenda un forte crescita degli utili in futuro, tale da giustificare un prezzo più alto[22]. Infatti un valore elevato del P/E *ratio* indica che gli investitori sono disposti a pagare un

[20] Si riveda quanto sottolineato riguardo al rapporto tra offerta di moneta e livelli dei prezzi secondo la Teoria Monetarista (Nota n. 2 a pag.1).
[21] Si riveda quanto sottolineato in *Introduzione* riguardo al ruolo delle aspettative (Nota n. 3 a pag.1).
[22] Si veda al proposito Ritter, J. R. (2005). *Economic Growth and Equity Returns. Pacific-Basin Finance Journal*, 13(5), 489-503 per la relazione tra aspettative di crescita e multipli di valutazione, incluso il P/E.

premio per un'azienda con prospettive di crescita superiori alla media, mentre un valore basso potrebbe suggerire aspettative di crescita modeste o rischi elevati.

6.1. Un primo approccio alla predizione

Se da un punto di vista interpretativo il modello selezionato, ottimizzato ed infine addestrato appare soddisfacente, un suo impiego predittivo, inteso come tentativo di anticipare di un certo numero di giorni il valore che l'indice DJIA assumerà, non ha dimostrato prestazioni al medesimo livello. Il modello minimo infatti mostra le seguenti prestazioni.

	RBF Mod.RMSE	RBF Mod._MAE	RBF Model_R2	Linear Mod._RMSE	Linear Model_MAE	Linear Model_R2	Polynomial Model_RMSE	Polynomial Model_MAE	Polynomial Model_R2	Linear Regr._RMSE	Linear Regr._MAE	Linear Regr._R2
0	0,057203	0,050973	0,946193	0,056948	0,050777	0,946672	0,045314	0,037143	0,966234	0,040526	0,030412	0,972993
1	0,057077	0,051026	0,948757	0,056058	0,050146	0,950571	0,044494	0,036177	0,96886	0,039998	0,030424	0,974835
2	0,057716	0,051449	0,946734	0,056289	0,050309	0,949335	0,04698	0,039284	0,964707	0,040228	0,030308	0,974123
3	0,056634	0,050594	0,946661	0,055557	0,049714	0,948669	0,047416	0,040417	0,96261	0,039429	0,02993	0,974146
4	0,0564	0,050207	0,949401	0,056305	0,050136	0,949572	0,044807	0,037092	0,968065	0,040722	0,030277	0,973622
5	0,057274	0,051063	0,948837	0,056913	0,05087	0,94948	0,044453	0,036142	0,969179	0,041145	0,030416	0,973595
6	0,056603	0,050626	0,948775	0,056739	0,050774	0,948529	0,051243	0,044065	0,958018	0,040265	0,030206	0,974078
7	0,056203	0,04999	0,94888	0,055038	0,048969	0,950977	0,045857	0,037573	0,965968	0,040462	0,030176	0,973504
8	0,056743	0,050584	0,948067	0,055321	0,049319	0,950638	0,04636	0,039151	0,965333	0,040717	0,03014	0,97326
9	0,056475	0,050411	0,949366	0,055473	0,049591	0,951146	0,046683	0,039314	0,965402	0,039963	0,030131	0,974646
Finale	0,056809	0,050666	0,948134	0,056473	0,050581	0,948744	0,047334	0,039928	0,963991	0,040263	0,03022	0,973947

Tabella 4: Risultati dei test sui vari modelli a 12 variabili – Predizione a 7 giorni

Esse appaiono piuttosto buone in termini di R^2. Analizziamo graficamente gli andamenti per confrontare l'accuratezza manifestata nei vari scenari. In Figura 18 abbiamo l'andamento globale sull'intero periodo. Esso conferma quanto descritto dal R^2, ossia la sostanziale analogia in termini di accuratezza del modello predittivo con quello interpretativo mostrato al paragrafo 6.3.

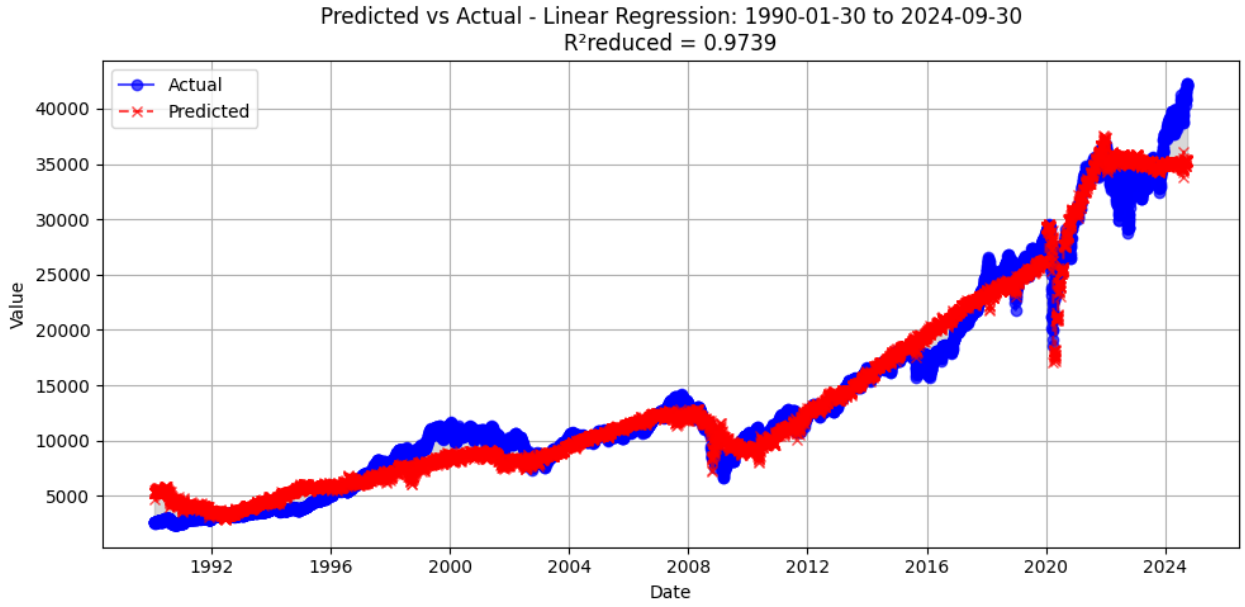


Figura 18: Confronto tra andamento predetto e andamento reale del DJIA – Modello predittivo a 7 giorni (12 var.)

Come si vede dalle Figure 19 e 20, gli scenari "*dot.com*" e "*11 settembre*" manifestano scostamenti apprezzabili tra gli andamenti predetti ed effettivi, come avveniva già per il modello interpretativo.

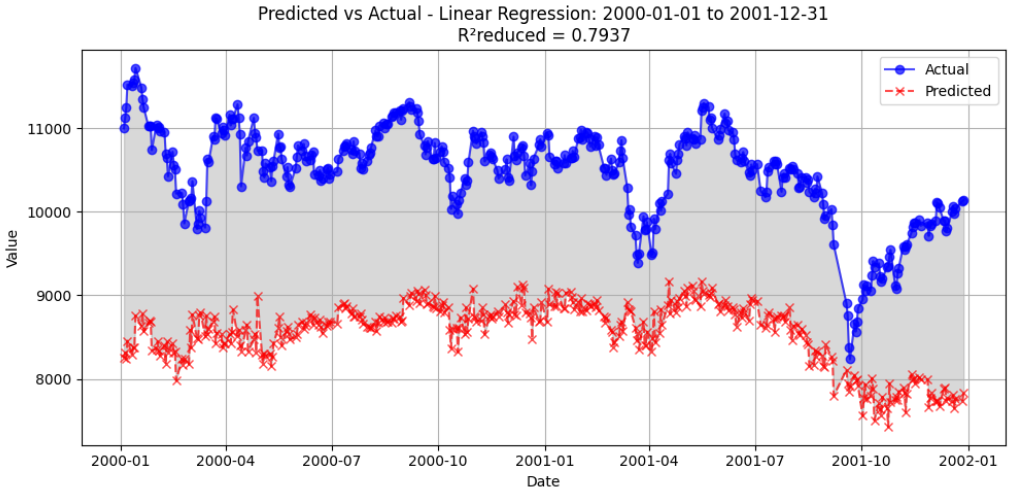


Figura 19: Confronto tra andamento predetto e andamento reale DJIA – Bolla delle dot.com – Modello predittivo a 7 giorni (12 var.)

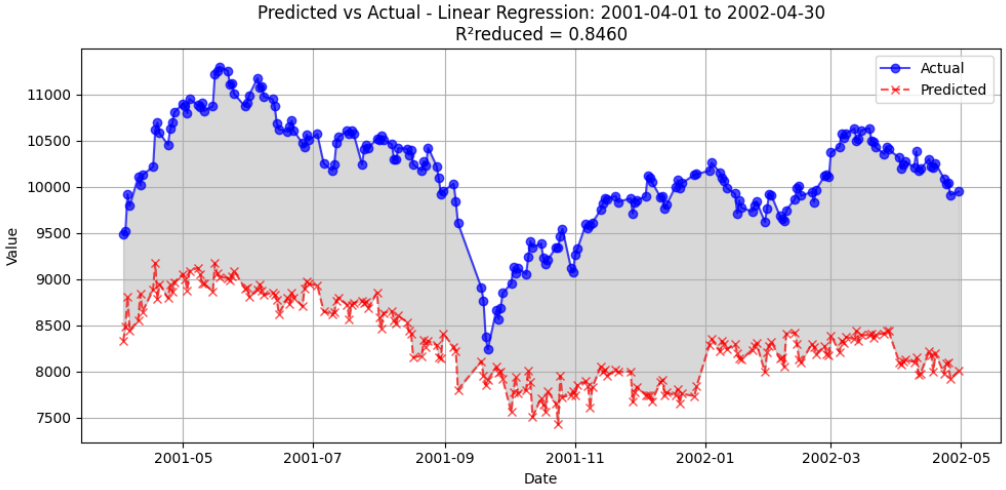


Figura 20: Confronto tra andamento predetto e andamento reale DJIA – Attacco alle Torri Gemelle (11 settembre) – Modello predittivo a 7 giorni (12 var.)

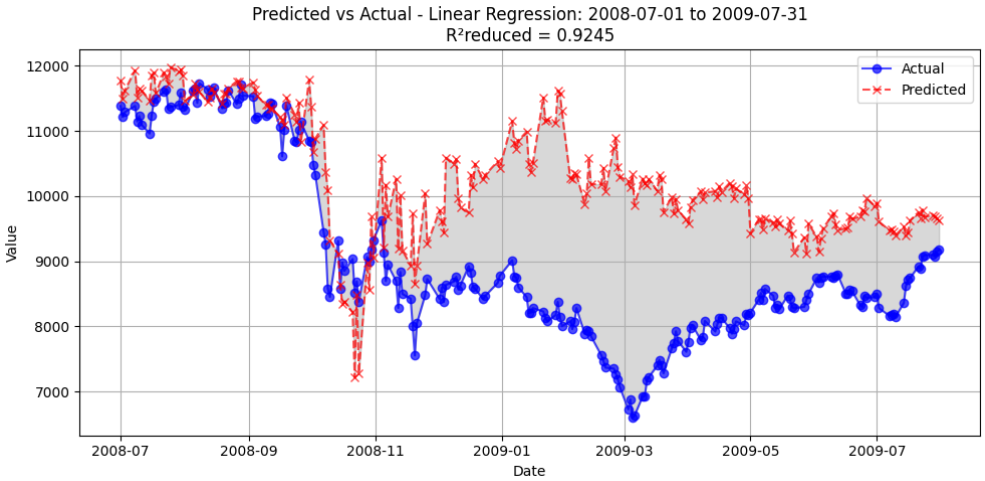


Figura 21: Confronto tra andamento predetto e andamento reale DJIA – Crisi Lehmann Brothers – Modello predittivo a 7 giorni (12 var.)

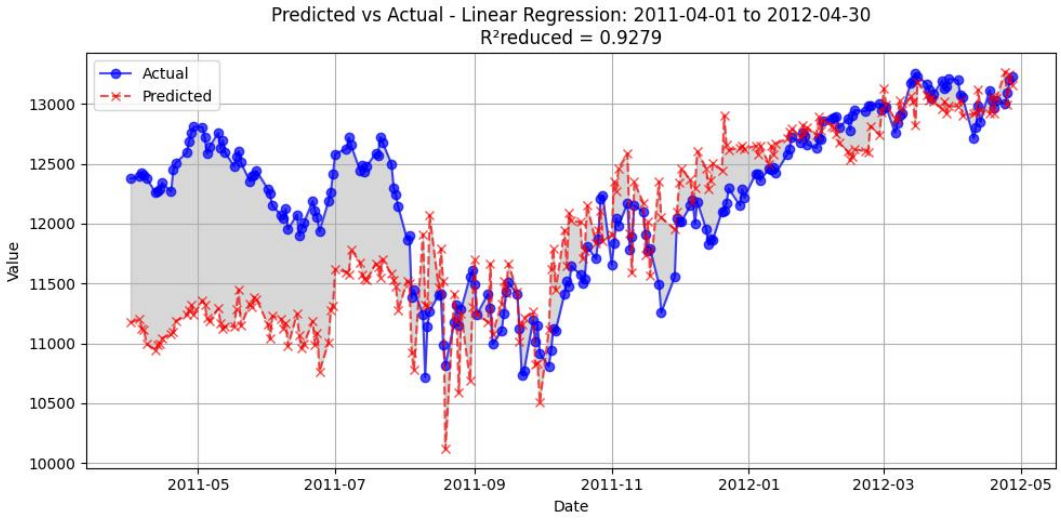


Figura 22: Confronto tra andamento predetto e andamento reale DJIA – Crisi del debito sovrano – Modello predittivo a 7 giorni (12 var.)

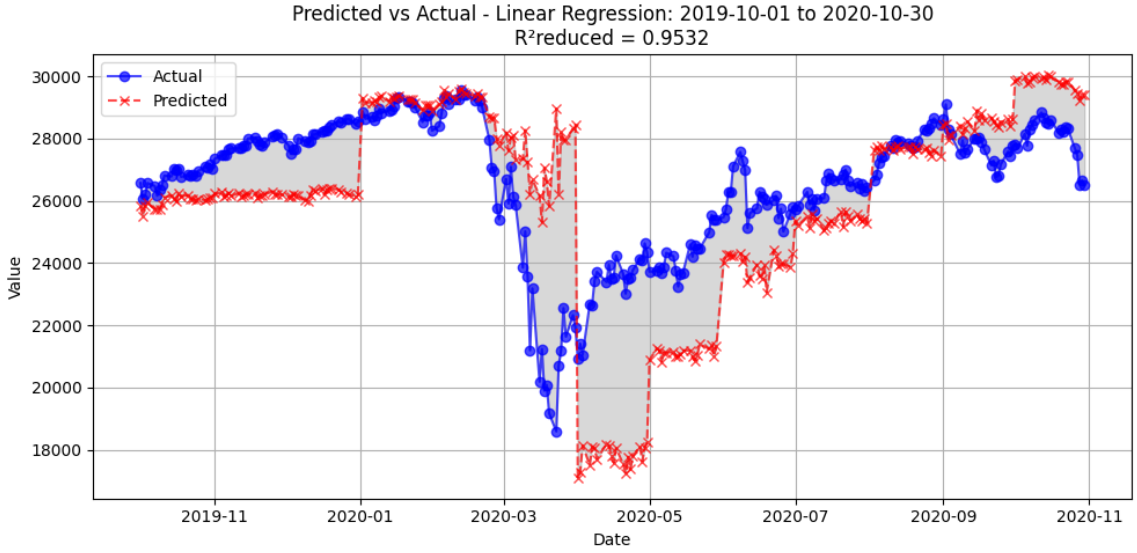


Figura 23: Confronto tra andamento predetto e andamento reale DJIA – Pandemia COVID-19 – Modello predittivo a 7 giorni (12 var.)

Anche negli altri tre scenari il modello predittivo si comporta analogamente a quello interpretativo.

Quanto visto per il modello a 12 variabili viene confermato per il modello a 32. Anche quest'ultimo esibisce un valore di R^2 piuttosto alto, solo lievemente inferiore a quello interpretativo.

	RBF Model_RMSE	RBF Model_MAE	RBF Model_R2	Linear Model_RMSE	Linear Model_MAE	Linear Model_R2	Polynomial Model_RMSE	Polynomial Model_MAE	Polynomial Model_R2	Linear Regr._RMSE	Linear Regr._MAE	Linear Regr._R2
0	0,051040423	0,044780308	0,957671523	0,047081681	0,040003032	0,963982959	0,054508916	0,047642189	0,951723119	0,032917596	0,024924538	0,982394015
1	0,050814942	0,044406206	0,959535597	0,047323362	0,040512873	0,964905308	0,054533015	0,047487626	0,953397493	0,03219307	0,024550924	0,983758915
2	0,051703865	0,045577265	0,956306775	0,04784697	0,040697529	0,96258231	0,054421893	0,047699885	0,951592195	0,03329345	0,02504279	0,981883025
3	0,05201262	0,0451027	0,954207926	0,049874191	0,04343568	0,957895881	0,055538606	0,048278573	0,947788906	0,032338685	0,024489797	0,982298195
4	0,052673184	0,046403953	0,955444293	0,048522345	0,041858495	0,962189905	0,055645728	0,048898413	0,950273504	0,032299684	0,024892163	0,983245913
5	0,052161259	0,045857521	0,956765644	0,048884378	0,042422793	0,962027164	0,055625455	0,048556812	0,950832286	0,032780989	0,024986709	0,982924361
6	0,051138175	0,044872897	0,958681151	0,047333071	0,040014741	0,964601314	0,054960611	0,048219001	0,952273359	0,033271394	0,025214835	0,982509599
7	0,05170204	0,045349673	0,955546148	0,047249378	0,039989304	0,96287331	0,055441777	0,048577088	0,948882651	0,032903659	0,025159444	0,981995443
8	0,051288804	0,044896742	0,958292517	0,04729368	0,040157229	0,964537033	0,055066331	0,048098077	0,951922583	0,033155956	0,025366781	0,982570218
9	0,052607692	0,046359295	0,954018342	0,048755877	0,04163296	0,960505185	0,055355853	0,048710436	0,949088813	0,032809891	0,025051283	0,982114721
Finale	0,052078088	0,045685564	0,956411829	0,047657628	0,040599971	0,963497432	0,055045288	0,048149057	0,951303374	0,032668285	0,02470801	0,982848139

Tabella 5: Risultati dei test sui vari modelli a 32 variabili – Predizione a 7 giorni

Tutti gli andamenti, sia quello complessivo che quelli in occasione delle grandi crisi confermano anch'essi, in linea generale, quanto visto per il modello interpretativo.

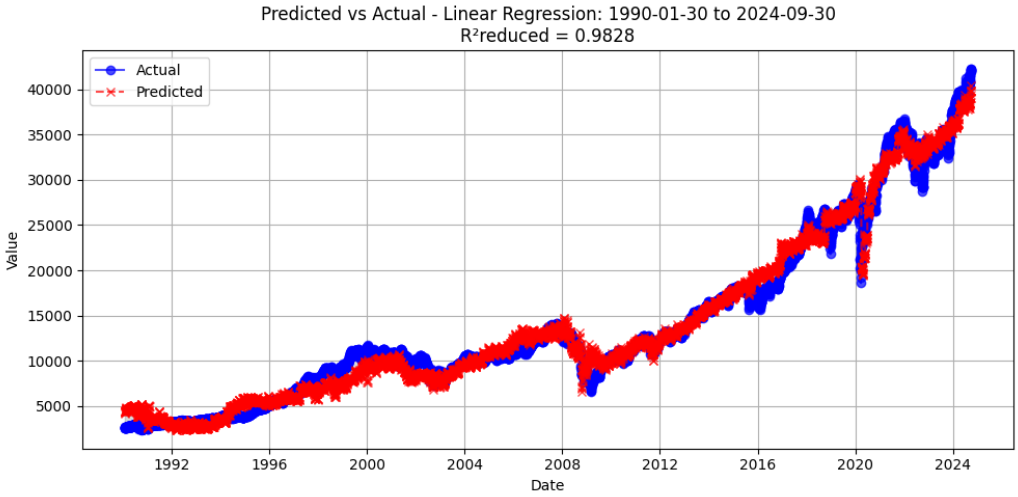


Figura 24: Confronto tra andamento predetto e andamento reale del DJIA – Modello predittivo a 7 giorni (32 var.)

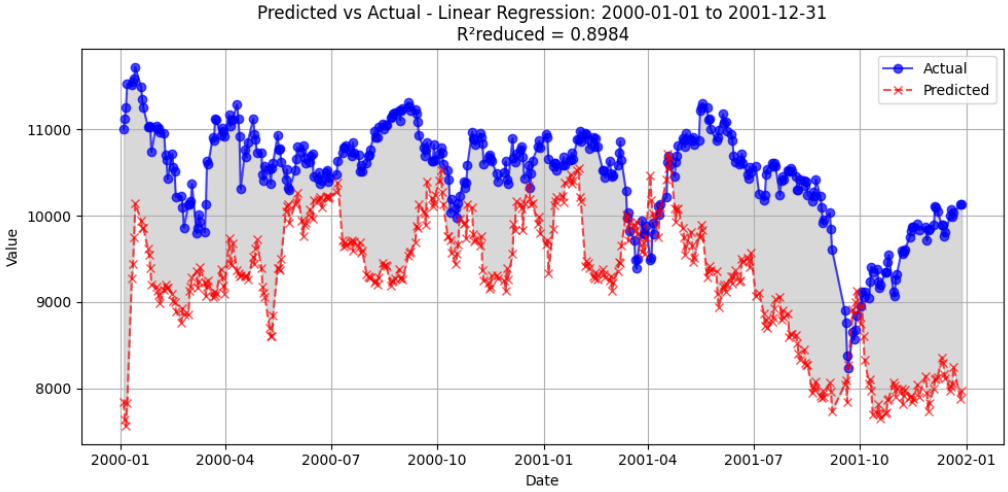


Figura 25: Confronto tra andamento predetto e andamento reale DJIA – Bolla delle dot.com – Modello predittivo a 7 giorni (32 var.)

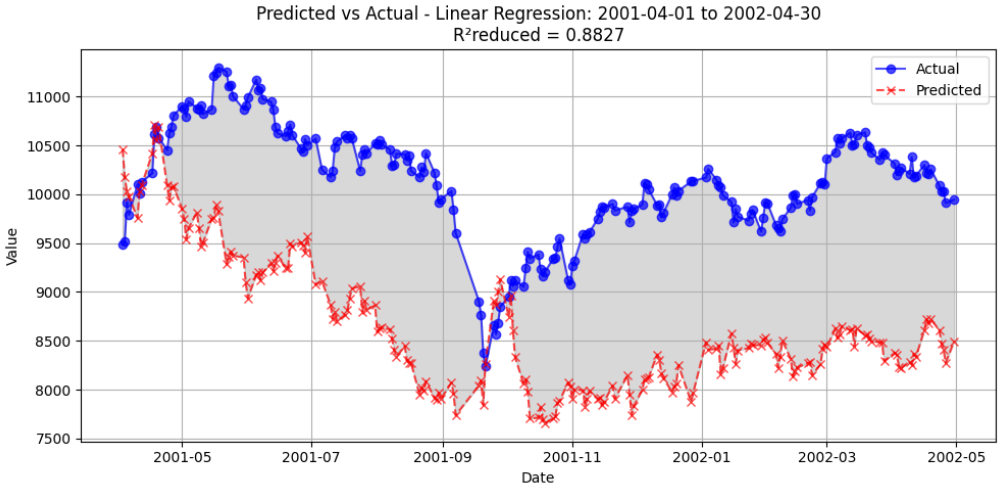


Figura 26: Confronto tra andamento predetto e andamento reale DJIA – Attacco alle Torri Gemelle (11 settembre) – Modello predittivo a 7 giorni (32 var.)

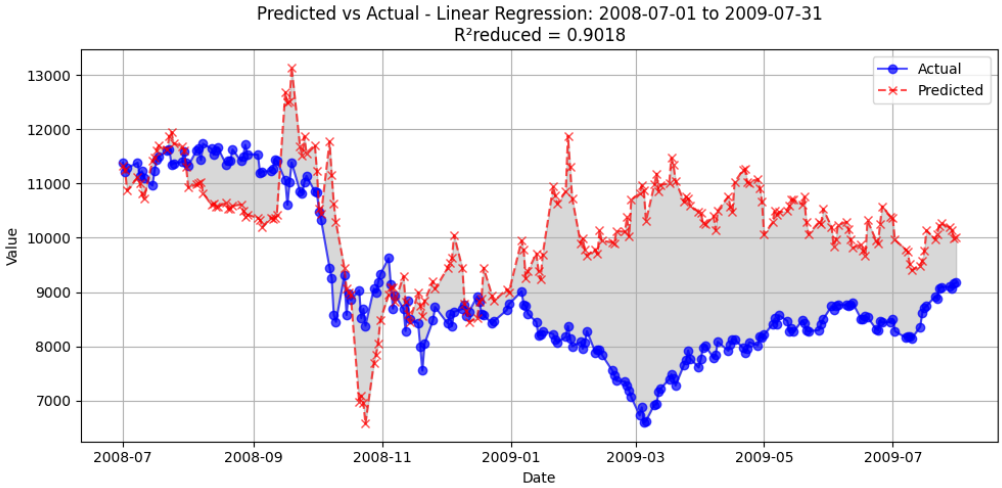


Figura 27: Confronto tra andamento predetto e andamento reale DJIA – Crisi Lehmann Brothers – Modello predittivo a 7 giorni (32 var.)

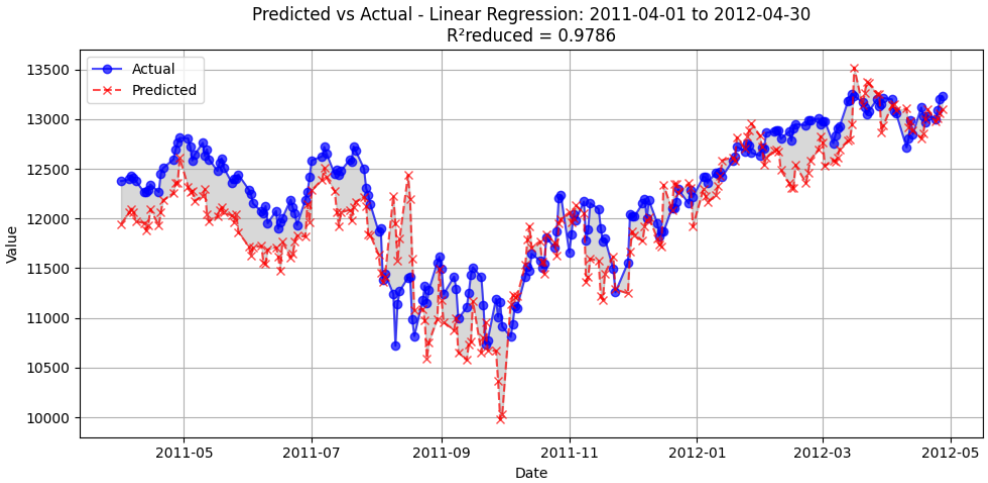


Figura 28: Confronto tra andamento predetto e andamento reale DJIA – Crisi del debito sovrano – Modello predittivo a 7 giorni (32 var.)

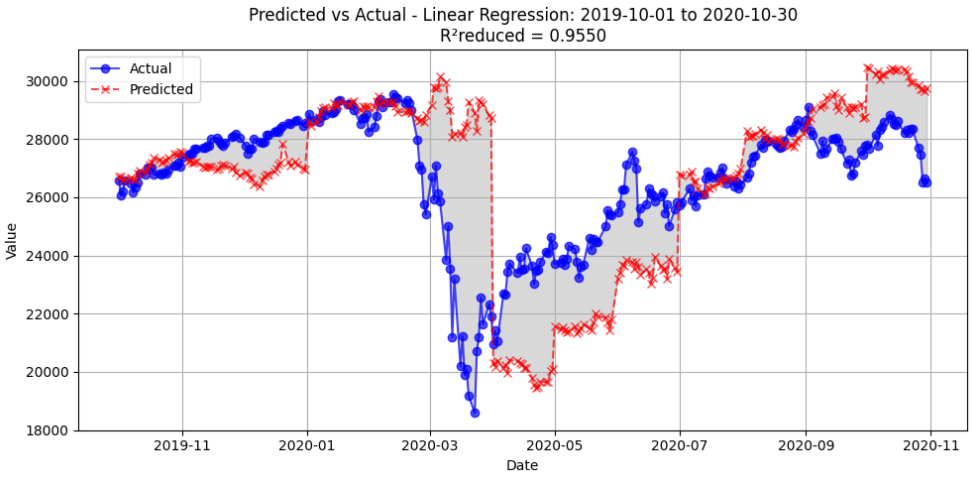


Figura 29: Confronto tra andamento predetto e andamento reale DJIA – Pandemia COVID-19 – Modello predittivo a 7 giorni (32 var.)

Le prestazioni appaiono comunque soddisfacenti, se rapportate a modelli interpretativi. Permangono le criticità di questi ultimi, manifestate in particolare negli scenari "*dot.com*" e "*11 settembre*" ed attribuibili a dinamiche non lineari non adeguatamente rappresentate dai modelli utilizzati.

6.2. Possibile miglioramento della predizione: Modello Autoregressivo

Allo scopo di incrementare ulteriormente l'accuratezza della previsione si è immaginato di introdurre alcuni attributi che possano rappresentare meglio quelle variabili non legate alla situazione generale macroeconomica, ma piuttosto alle valutazioni degli operatori economici ed alle loro aspettative o considerazioni sull'andamento dei prezzi. A tal proposito sono stati inclusi nel modello i seguenti attributi:

- "*MME_Prices*": medie mobili esponenziali calcolate sulla colonna "*Price*".
- "*MME_vs_Prices*": differenza ponderata (tramite il parametro "*MMEPricesWeightModel*") tra il valore del prezzo e la media mobile;
- "*PriceVariation*": variazione marginale dei prezzi, calcolata a partire dalle istanze giornaliere degli stessi;

allo scopo di valutare le tendenze di fondo dei prezzi e confrontarle con gli andamenti giornalieri puntuali, ed infine le istanze ritardate fino a 7 giorni della variabile dipendente ("*Price*") andando di fatto a implementare un modello autoregressivo.

	RBF _RMSE	RBF _MAE	RBF _R2	Linear _RMSE	Linear _MAE	Linear _R2	Polyno-mial _RMSE	Polyno-mial _MAE	Polyno-mial _R2	Linear Regr_RMSE	Linear Regr_MAE	Linear Regr_R2
0	0,049823	0,039973	0,960636	0,055956	0,050352	0,950349	0,054038	0,044939	0,953695	0,011678	0,00724	0,997837
1	0,050244	0,040847	0,959298	0,059038	0,053649	0,943804	0,053728	0,044941	0,953458	0,011524	0,007158	0,997859
2	0,049626	0,040024	0,959566	0,056521	0,050938	0,947549	0,05381	0,044691	0,95246	0,011363	0,007091	0,99788
3	0,048615	0,038685	0,963038	0,055246	0,049546	0,952267	0,053491	0,044392	0,955252	0,011739	0,00754	0,997845
4	0,049435	0,039482	0,960808	0,056252	0,050667	0,949254	0,053931	0,044826	0,953355	0,0119	0,007063	0,997729
5	0,050325	0,040763	0,959222	0,056058	0,050325	0,949402	0,054268	0,04515	0,952581	0,01171	0,006961	0,997792
6	0,049843	0,039751	0,958971	0,060147	0,055042	0,940254	0,053903	0,044636	0,952014	0,011813	0,007274	0,997695
7	0,050306	0,040557	0,959552	0,058684	0,053303	0,944958	0,05409	0,044929	0,953238	0,011653	0,007224	0,99783
8	0,049495	0,039681	0,96034	0,056247	0,050583	0,948781	0,054111	0,045081	0,952598	0,0117	0,007177	0,997784
9	0,048476	0,038603	0,962053	0,056769	0,051416	0,94796	0,052888	0,043773	0,954832	0,011578	0,00717	0,997835
Finale	0,049291	0,039368	0,960806	0,05625	0,050583	0,948956	0,053619	0,04461	0,95362	0,011486	0,007208	0,997872

Tabella 4: Parametri del modello minimo (a 12 variabili) ed interpretazione

I parametri migliorano apprezzabilmente, ed anche in questo caso il modello di regressione lineare risulta il migliore. Di seguito gli andamenti predetti da quest'ultimo, sia nello scenario generale, che nelle 5 crisi.

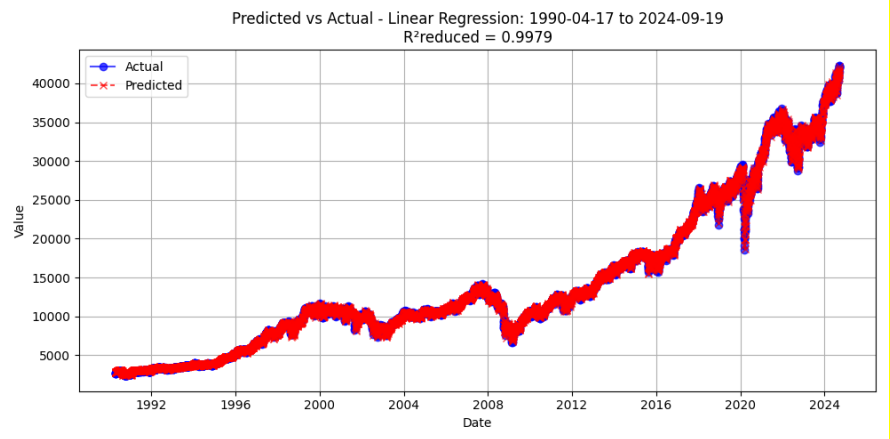


Figura 30: Predizione a 7 giorni – Modello autoregressivo

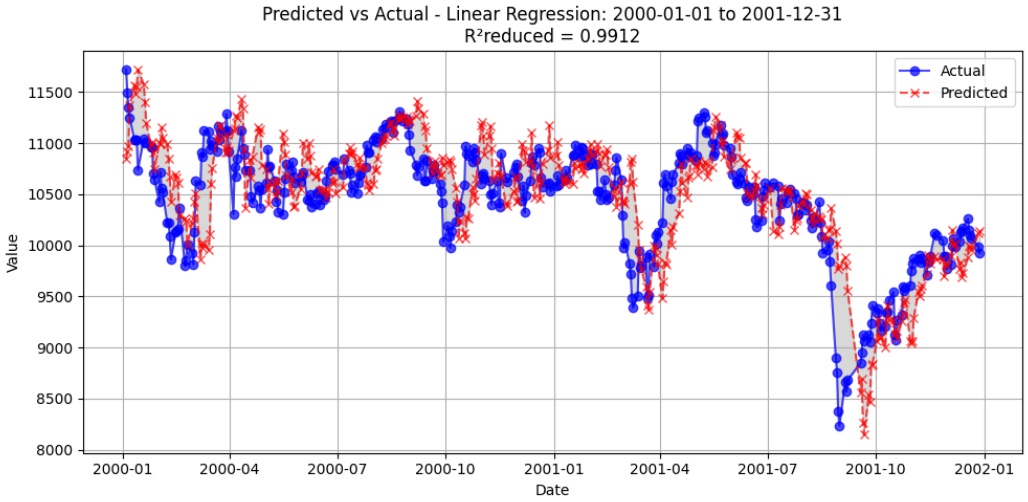


Figura 31: Predizione a 7 giorni – Bolla delle dot.com – Modello autoregressivo

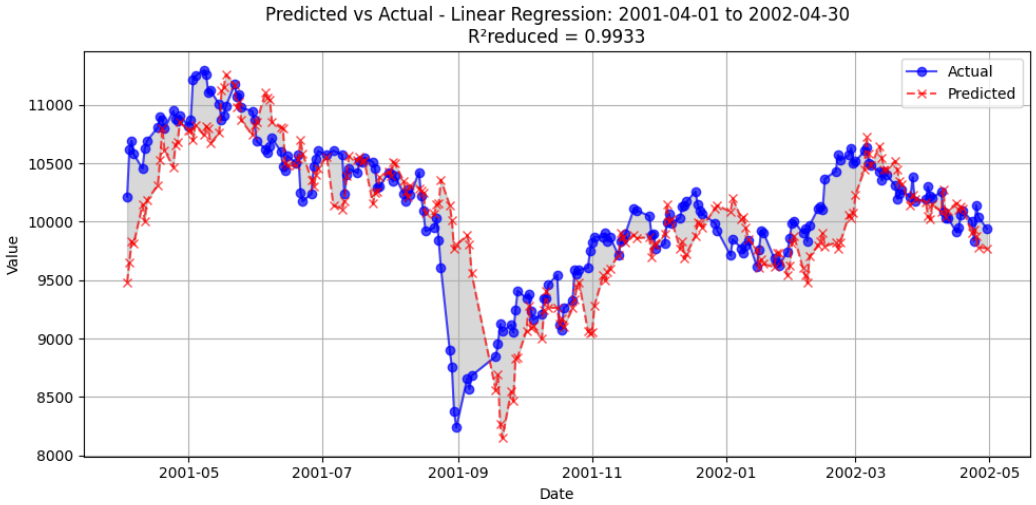


Figura 32: Predizione a 7 giorni – Attacco alle Torri Gemelle (11 settembre) – Modello autoregressivo

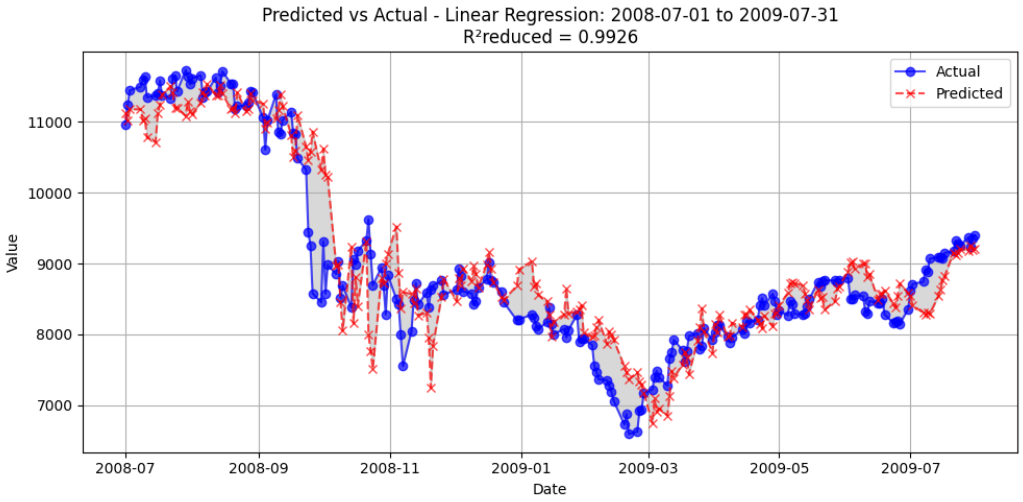


Figura 33: Predizione a 7 giorni – Crisi Lehmann Brothers – Modello autoregressivo

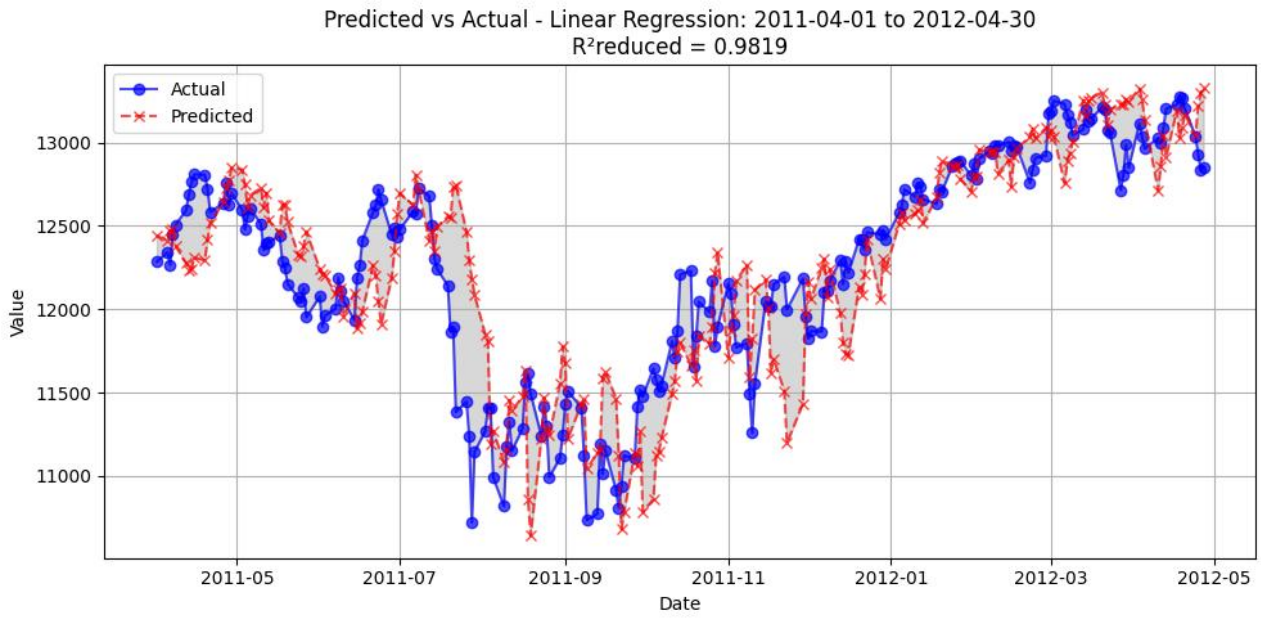


Figura 34: Predizione a 7 giorni – Crisi del debito sovrano – Modello autoregressivo

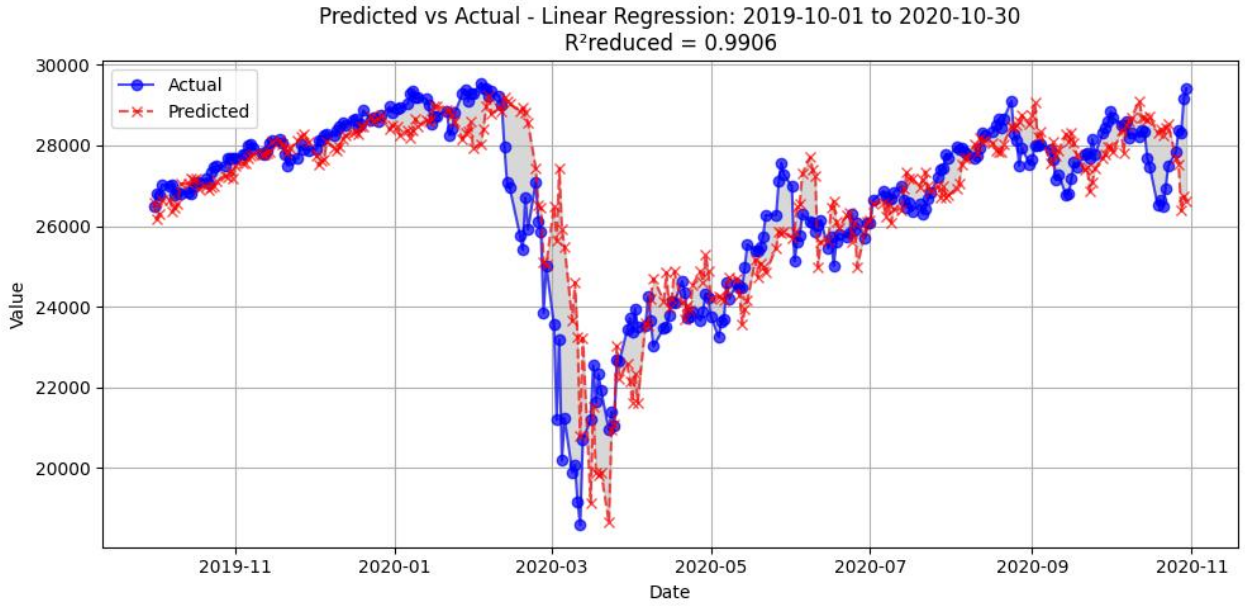


Figura 35: Predizione a 7 giorni – Pandemia COVID-19 – Modello autoregressivo

Il modello sembra molto accurato, anche se l'andamento predetto appare una copia ritardata dell'andamento effettivo. Sulla base di tale osservazione, nel prossimo paragrafo emergeranno delle considerazioni che dimostreranno come il modello predittivo autoregressivo proposto non sia accettabile.

6.3. Approfondimenti sul modello autoregressivo

L'analisi di correlazione condotta sul modello di cui al precedente paragrafo, e i cui elementi salienti sono mostrati nelle seguenti Figura 36 e Tabella 5 consente di fare immediatamente alcune osservazioni.

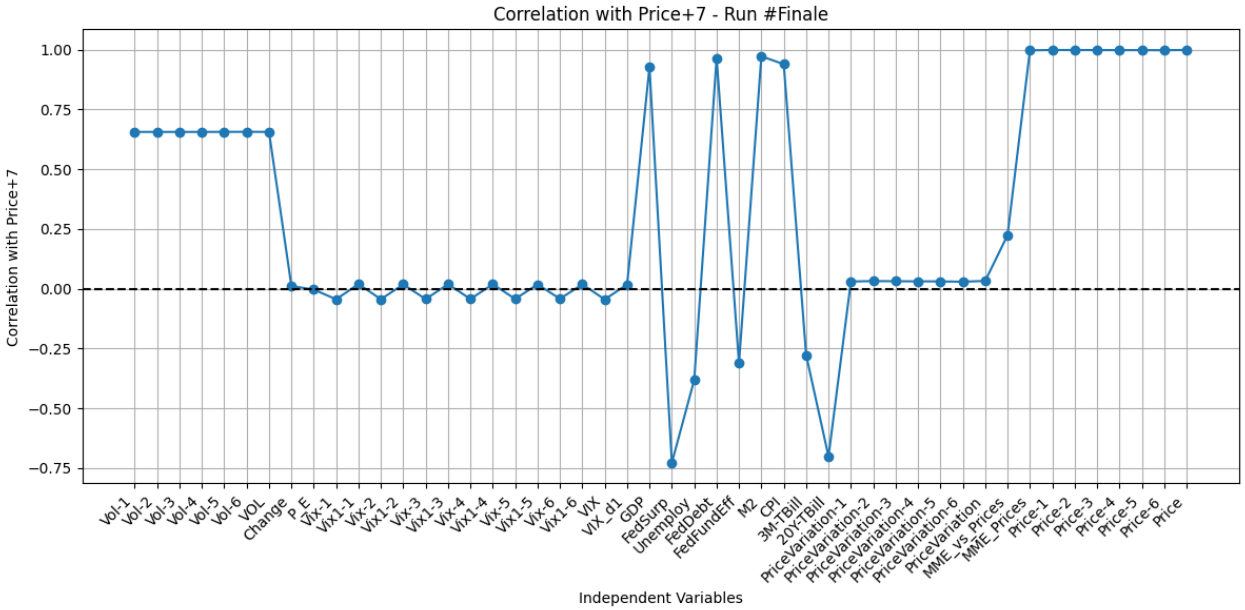


Figura 36: Correlation Line modello autoregressivo

	Vol-1	….	MME_vs_*Prices*	MME_Prices	Price-1	Price-2	Price-3	Price-4	Price-5	Price-6	Price
Vol-1	1	….	-0,08945	0,669066	0,655555	0,656613	0,65774	0,658848	0,65966	0,660646	0,65546
….	….	….	….	….	….	….	….	….	….	….	….
MME_Prices	0,669066	….	0,168852	1	0,998295	0,998403	0,998501	0,998591	0,998672	0,998746	0,998179
Price-1	0,655555	….	0,223216	0,998295	1	0,999822	0,999675	0,999506	0,999342	0,999192	0,999822
Price-2	0,656613	….	0,219077	0,998403	0,999822	1	0,999822	0,999675	0,999506	0,999341	0,999675
Price-3	0,65774	….	0,214743	0,998501	0,999675	0,999822	1	0,999822	0,999674	0,999506	0,999507
Price-4	0,658848	….	0,210606	0,998591	0,999506	0,999675	0,999822	1	0,999822	0,999674	0,999342
Price-5	0,65966	….	0,206836	0,998672	0,999342	0,999506	0,999674	0,999822	1	0,999822	0,999192
Price-6	0,660646	….	0,203067	0,998746	0,999192	0,999341	0,999506	0,999674	0,999822	1	0,999034
Price	0,65546	….	0,228	0,998179	0,999822	0,999675	0,999507	0,999342	0,999192	0,999034	1

Tabella 5: Tabella correlazioni modello autoregressivo: prezzi e relative istanze precedenti

La variabile "*Price*" (variabile dipendente) ha una correlazione molto prossima a 1 con tutte le sue istanze precedenti.

L'applicazione della metodologia di ottimizzazione vista al Capitolo 6 consente di identificare i seguenti parametri base:

- $M = 100$;
- $m = 48$;
- $n = 7669$ (Numero delle Osservazioni);
- $y = Price + 7$ (Predizione della variabile "*Price+7*");

Ancora una volta il modello per l'ottimizzazione è stato implementato 3 volte assegnando i seguenti valori massimi per il coefficiente di correlazione a coppie:

- $C_{max} = 0{,}7$;
- $C_{max} = 0{,}5$;
- $C_{max} = 0{,}3$.

I risultati conseguiti sono esposti di seguito.

Caso C_{max}=0,7

- Numero di variabili indipendenti: $k = 22$;
- $R^2 \approx 1$.

Risultati ottenuti:

Variabili Indipendenti	$\boldsymbol{\beta}$	$\boldsymbol{z}$
Vol-5	0	1
P_E	0	1
Vix1-1	0	1
Vix1-2	0	1
Vix1-3	0	1
Vix1-4	0	1
Vix1-5	0	1
Vix1-6	0	1
CBOE Volatility Index: VIX	0	1
CBOE Volatility Index: VIX_1st_derivative	0	1
Federal Surplus or Deficit [-]	0	1
Unemployment Rate	0	1
Federal Funds Effective Rate	0	1
PriceVariation-1	0	1
PriceVariation-2	0	1
PriceVariation-3	0	1
PriceVariation-4	0	1
PriceVariation-5	0	1
PriceVariation-6	0	1
PriceVariation	0	1
MME_vs_Prices	0	1
Price	1	1
β_0 *(Intercetta)*	0	

Caso C_{max}=0,5

- Numero di variabili indipendenti: $k = 21$;
- $R^2 \approx 1$.

Risultati ottenuti:

Variabili Indipendenti	$\boldsymbol{\beta}$	$\boldsymbol{z}$
P_E	0	1
Vix1-1	0	1
Vix1-2	0	1
Vix1-3	0	1
Vix1-4	0	1
Vix1-5	0	1
Vix1-6	0	1
CBOE Volatility Index: VIX	0	1
CBOE Volatility Index: VIX_1st_derivative	0	1
Federal Surplus or Deficit [-]	0	1
Unemployment Rate	0	1
Federal Funds Effective Rate	0	1

Variabili Indipendenti	$\boldsymbol{\beta}$	$\boldsymbol{z}$
PriceVariation-1	0	1
PriceVariation-2	0	1
PriceVariation-3	0	1
PriceVariation-4	0	1
PriceVariation-5	0	1
PriceVariation-6	0	1
PriceVariation	0	1
MME_vs_Prices	0	1
Price	1	1
β_0 *(Intercetta)*	0	

Caso C_{max}=0,3

- Numero di variabili indipendenti: $k = 19$;
- $R^2 \approx 1$.

Risultati ottenuti:

Variabili Indipendenti	$\boldsymbol{\beta}$	$\boldsymbol{z}$
P_E	0	1
Vix1-1	0	1
Vix1-2	0	1
Vix1-3	0	1
Vix1-4	0	1
Vix1-5	0	1
Vix1-6	0	1
CBOE Volatility Index: VIX_1st_derivative	0	1
Unemployment Rate	0	1
Federal Funds Effective Rate	0	1
PriceVariation-1	0	1
PriceVariation-2	0	1
PriceVariation-3	0	1
PriceVariation-4	0	1
PriceVariation-5	0	1
PriceVariation-6	0	1
PriceVariation	0	1
MME_vs_Prices	0	1
Price	1	1
β_0 *(Intercetta)*	0	

I risultati ottenuti in precedenza mettono in risalto che, benché sia stata soddisfatta la relazione:

$$\begin{cases} z_i = 0 \; se \; \beta_i = 0 \\ z_i = 1 \; se \; \beta_i \neq 0 \end{cases}$$

ad eccezione della variabile "*Price*", tutte le altre variabili individuate dal modello sono caratterizzate da valori di $\beta_i \neq 0$ talmente piccoli da non dare alcun contributo.

Ciò esprime il fatto che la correlazione tra la variabile dipendente e la sua istanza precedente più prossima – presa come variabile indipendente – è così forte da mettere in ombra tutti gli altri contributi, rendendo di

fatto inconsistente ed inutile il modello stesso, che all'atto pratico non fa altro che replicare con un ritardo di 7 giorni l'andamento della variabile "*Price*" di cui dispone, conseguendo comunque ottimi risultati in termini di R^2 in virtù della ridotta devianza della grandezza "*Price*" su orizzonti limitati nel tempo, tranne che in periodi particolari come quelli delle grandi crisi, dove l'andamento anomalo della grandezza predetta rispetto a quella effettiva appare particolarmente evidente.

6.4. Un modello alternativo per la predizione: *Random Forest Regression*

Visti gli esiti di cui al paragrafo precedente, ed allo scopo comunque di migliorare la predizione e interpretare meglio le non linearità insite nelle dinamiche effettive, si è deciso di aggiungere ai precedenti un modello basato sulla *random forest regression* che, auspicabilmente, migliori la capacità del modello di estrarre dai dati ciò che condiziona l'immediato futuro.

Utilizzando i 12 attributi del modello minimo (NON autoregressivo) le *performance* che si ottengono con un orizzonte previsionale di 7 giorni sono mostrate di seguito.

	RandomForest_MSE	RandomForest_RMSE	RandomForest_MAE	RandomForest_R2
0	0,000125041	0,011182175	0,00621871	0,997940434
1	0,000109237	0,010451627	0,006176025	0,998203184
2	0,000123571	0,011116248	0,006256205	0,997990813
3	0,000128087	0,011317556	0,006175313	0,997948764
4	0,000121812	0,01103686	0,006352721	0,998084314
5	0,000136316	0,011675444	0,006455254	0,997853155
6	0,000105428	0,010267804	0,006168703	0,998280507
7	0,000117497	0,0108396	0,006010111	0,998137181
8	0,000129132	0,011363631	0,006517215	0,997983736
9	0,000114172	0,010685151	0,00611742	0,998158166
Finale	9,59776E-05	0,009796818	0,005315584	0,998456803

Tabella 6: Risultati dei test sul modello a 12 variabili – Predizione a 7 giorni – RANDOM FOREST REGRESSION

Tutte le metriche esibiscono valori ottimi, ma anche alla luce di quanto visto per gli altri modelli, è necessario analizzare graficamente gli andamenti per verificarne l'adattamento ai dati reali.

Le Figure seguenti (37-38-39-40-41-42) raffrontano gli andamenti predetti con quelli effettivi.

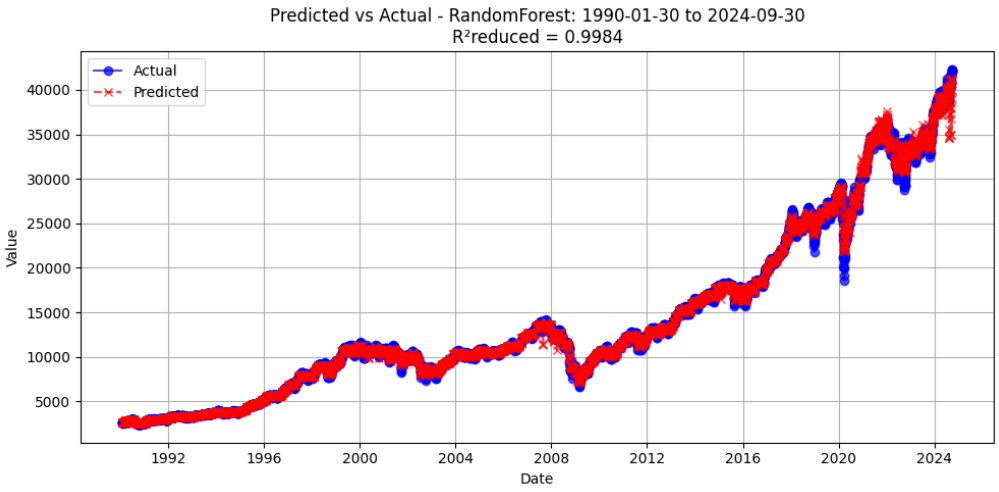


Figura 37: Random forest (12 variabili) – Intero periodo

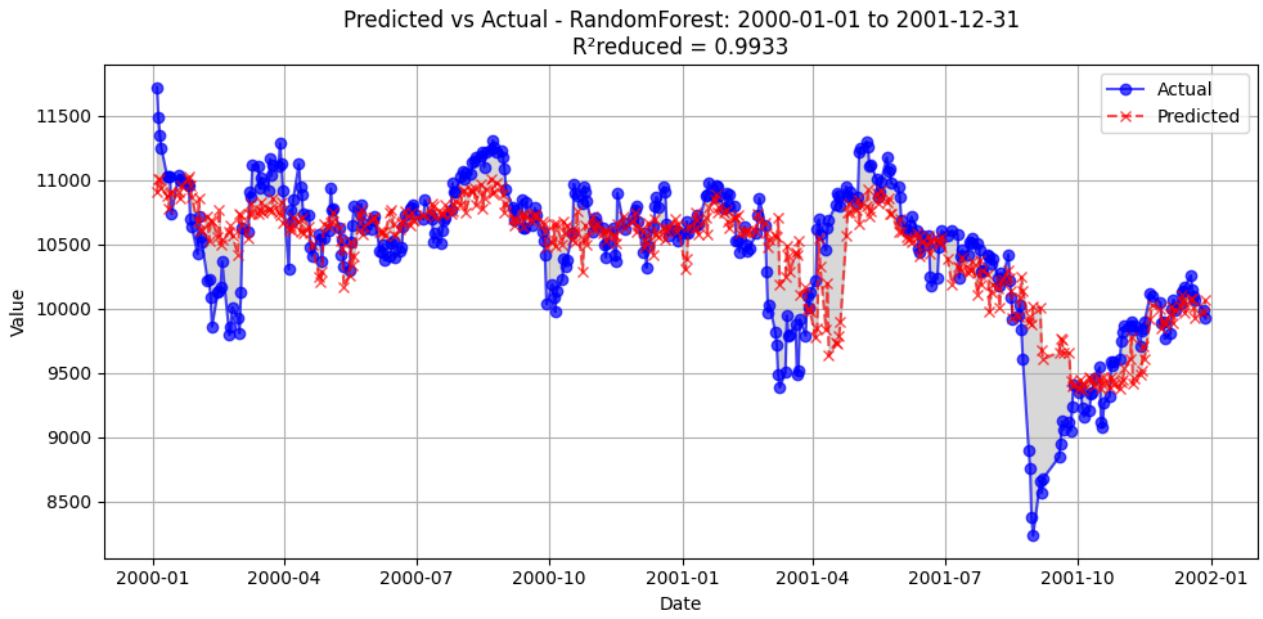


Figura 38: Random forest (12 variabili) – Bolla delle dot.com

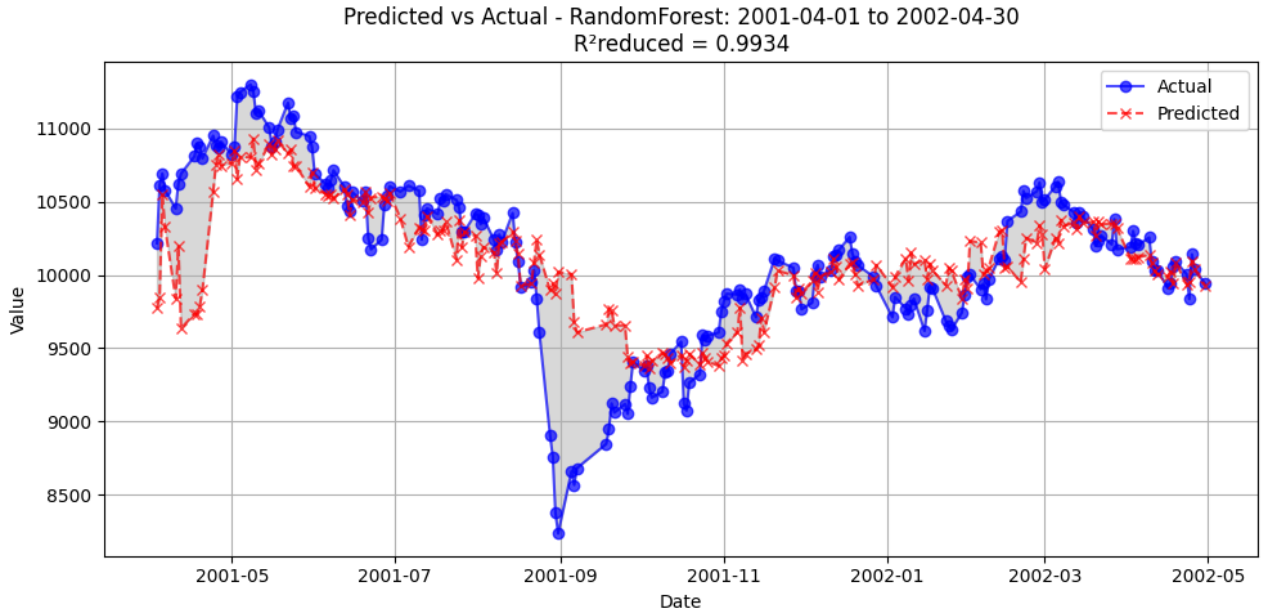


Figura 39: Random forest (12 variabili) – Attacco alle Torri Gemelle (11 settembre)

Le crisi "*dot.com*" e "*11 settembre*" costituivano il principale tallone di Achille dei modelli basati sulla regressione lineare. Il modello basato sulla *random forest* sfoggia prestazioni ottime in tali situazioni, mostrando tuttavia la tendenza a "tagliare" i crolli più repentini, il che è tutto sommato comprensibile, poiché, soprattutto nel caso degli attacchi alle Torri Gemelle, il movente principale del crollo è un evento *esterno* alla galassia dei parametri che vengono considerati nel modello e pertanto quest'ultimo non ha modo di percepire istantaneamente la sua influenza sulle scelte degli investitori. La sua percezione avviene in un secondo momento, tramite le variazioni dei parametri del mercato. Si può notare infatti come l'accuratezza nella fase di reazione al crollo sia decisamente elevata.

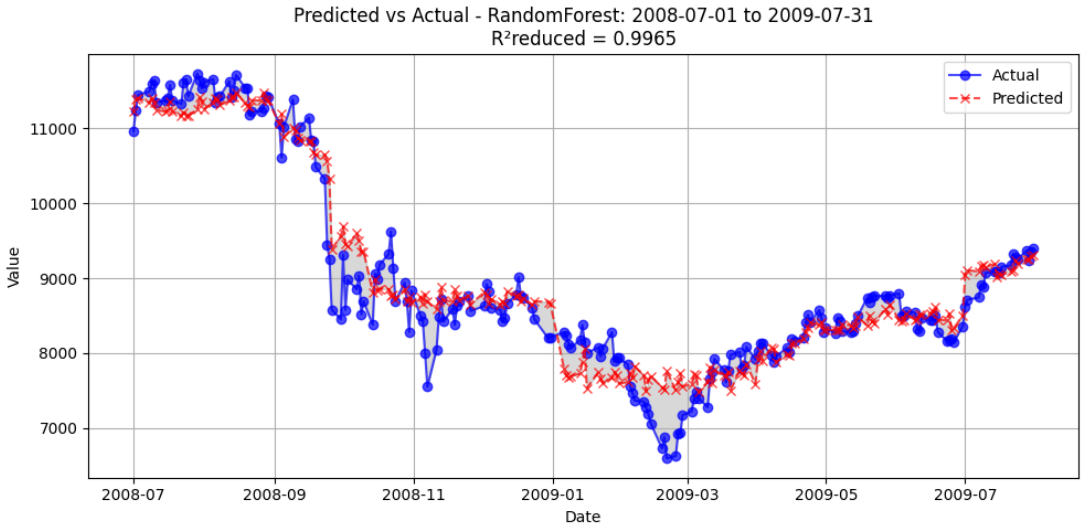


Figura 40: Random forest (12 variabili) – Crisi Lehmann Brothers

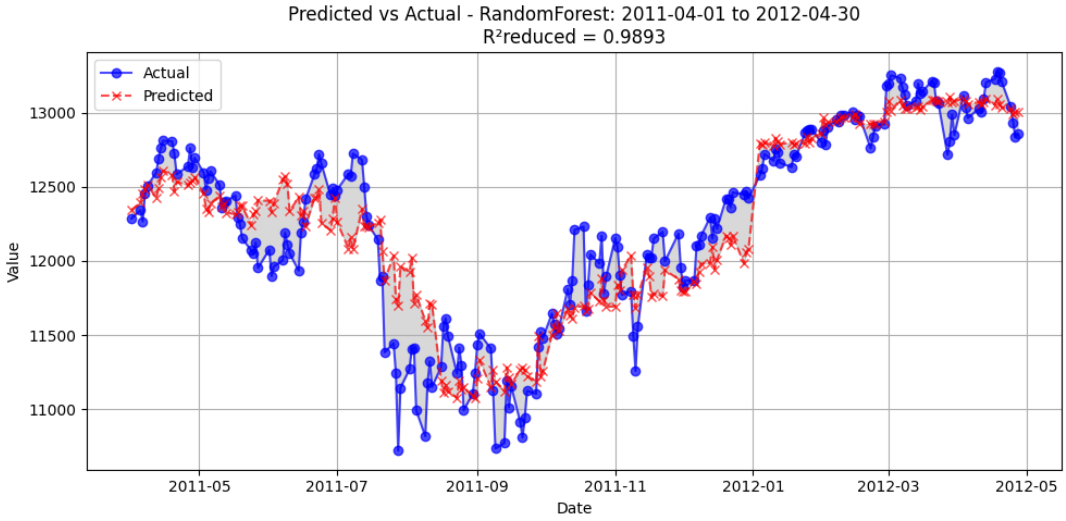


Figura 41: Random forest (12 variabili) – Crisi del debito sovrano

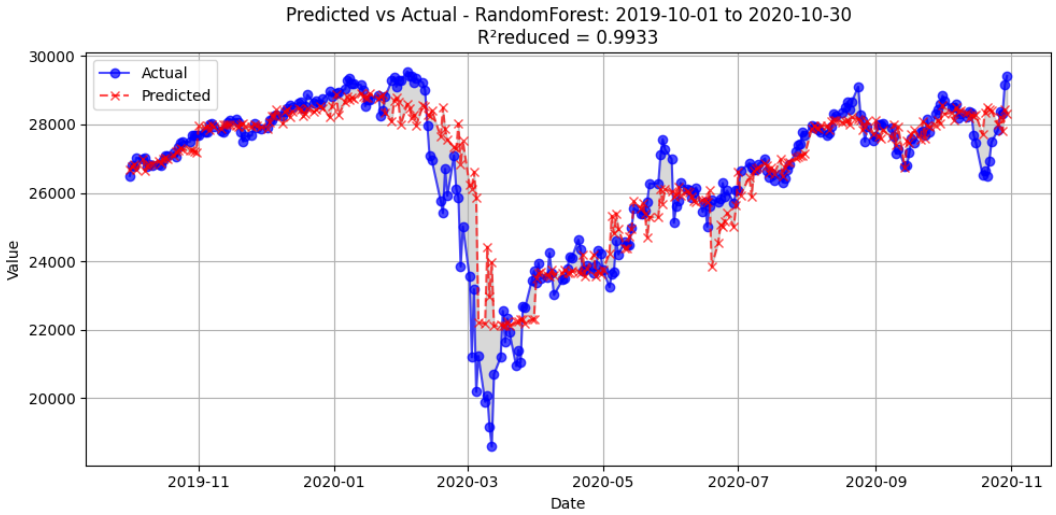


Figura 42: Random forest (12 variabili) – Pandemia COVID-19

In definitiva, l'accuratezza è migliorata in maniera considerevole, soprattutto nelle grandi crisi. Il tutto è testimoniato anche dai diagrammi dei residui, mostrati nella Figura 43, che esibiscono un andamento migliore anche dei modelli interpretativi, a conferma del fatto che la *random forest* cattura sicuramente meglio le dinamiche non lineari rispetto alla regressione lineare.

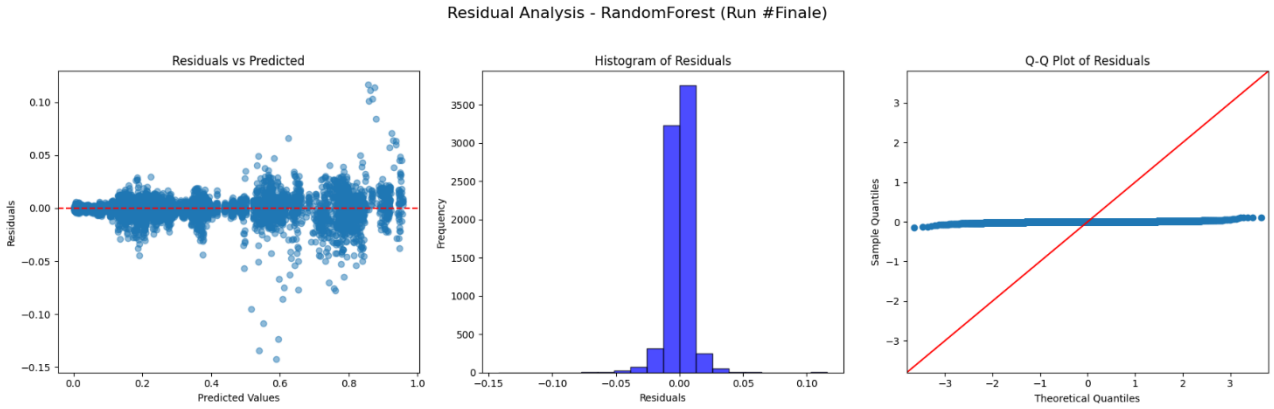


Figura 43: Random forest (12 variabili) – Pandemia COVID-19

Utilizzando il modello a 32 variabili, inoltre, l'accuratezza, generalmente, migliora ancora[23].

	RandomForest_MSE	RandomForest_RMSE	RandomForest_MAE	RandomForest_R2
0	7,54486E-05	0,008686113	0,005112104	0,99879571
1	6,41808E-05	0,008011293	0,004892535	0,998914293
2	6,79846E-05	0,008245278	0,00498518	0,998898716
3	6,97848E-05	0,008353732	0,00506871	0,998890462
4	6,27284E-05	0,007920124	0,00489805	0,998943246
5	6,85847E-05	0,008281586	0,004963141	0,998905005
6	7,41807E-05	0,008612819	0,005107317	0,998819084
7	6,39044E-05	0,00799402	0,005039496	0,998973281
8	7,04786E-05	0,008395155	0,005042289	0,99886544
9	7,67558E-05	0,008761039	0,005290185	0,998763151
Finale	5,05404E-05	0,007109174	0,004162579	0,999187376

Tabella 7: Risultati dei test sul modello a 32 variabili – Predizione a 7 giorni – RANDOM FOREST REGRESSION

In particolare, il MSE si dimezza passando dalle 12 alle 32 variabili.

Anche i grafici dei residui testimoniano il miglioramento dell'accuratezza.

[23] Quanto sperimentato è in linea con le conclusioni di altri studi analoghi. Si veda al proposito Kumar, I., Dogra, K., Utreja, C., & Yadav, P., 2018. *A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction*. 2018 *Second International Conference on Inventive Communication and Computational Technologies* (ICICCT), pp. 1003-1007:

"….. *The experimental results show that Random Forest algorithm performs the best for large datasets and Naive Bayesian Classifier is the best for small datasets. The results also reveal that reduction in the number of technical indicators reduces the accuracies of each algorithm…*."

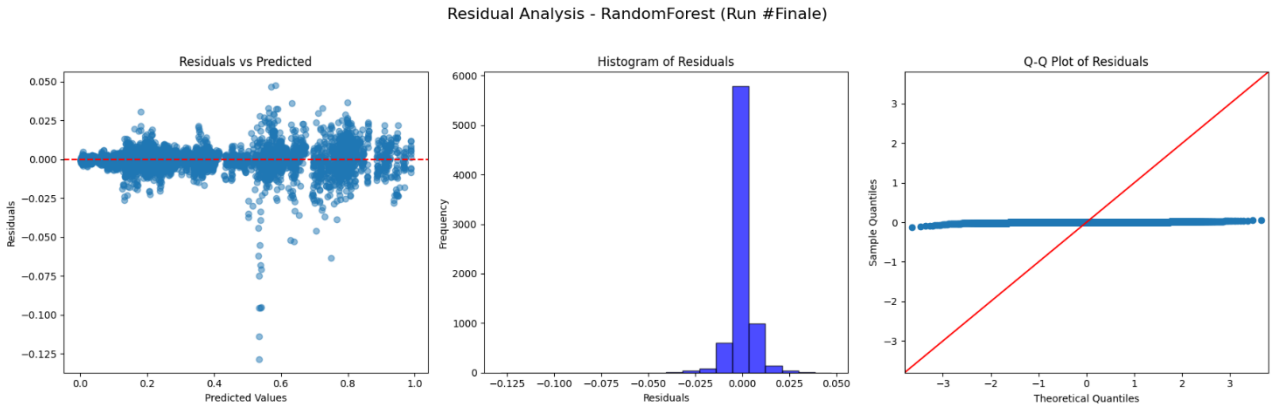


Figura 44: Random forest (12 variabili) – Pandemia COVID-19

Di seguito, ancora una volta, gli andamenti predetti raffrontati con quelli effettivi.

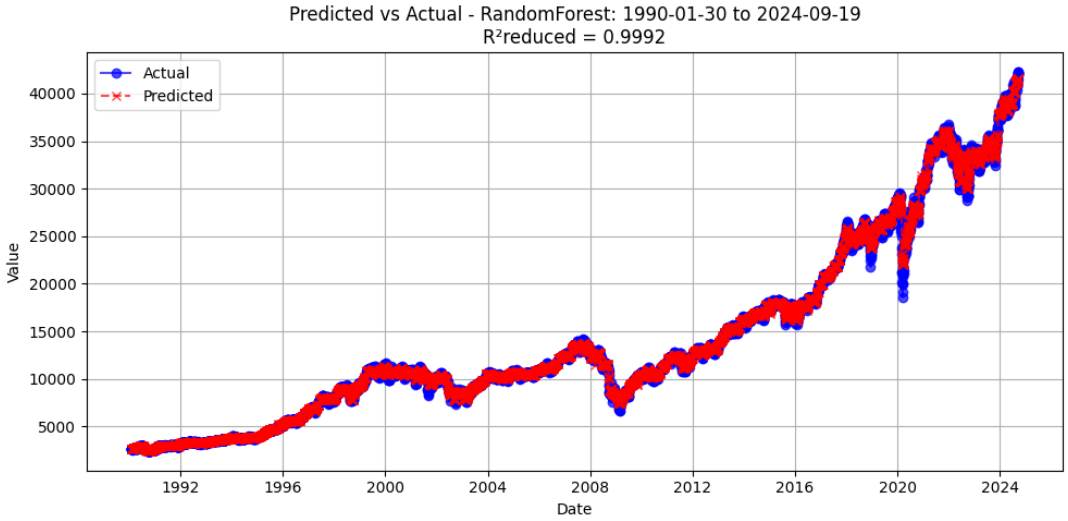


Figura 45: Random forest (32 variabili) – Intero periodo

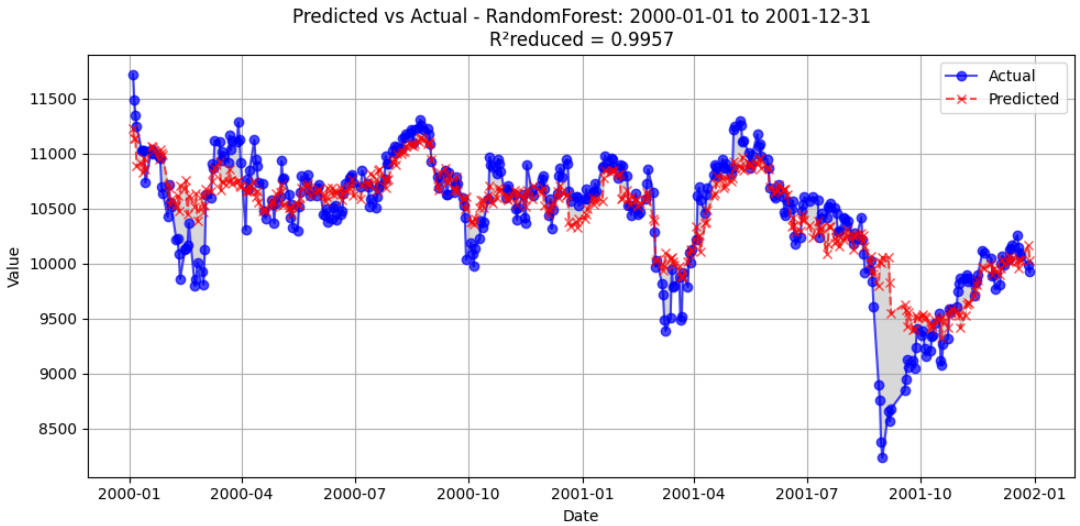


Figura 46: Random forest (32 variabili) – Bolla delle dot.com

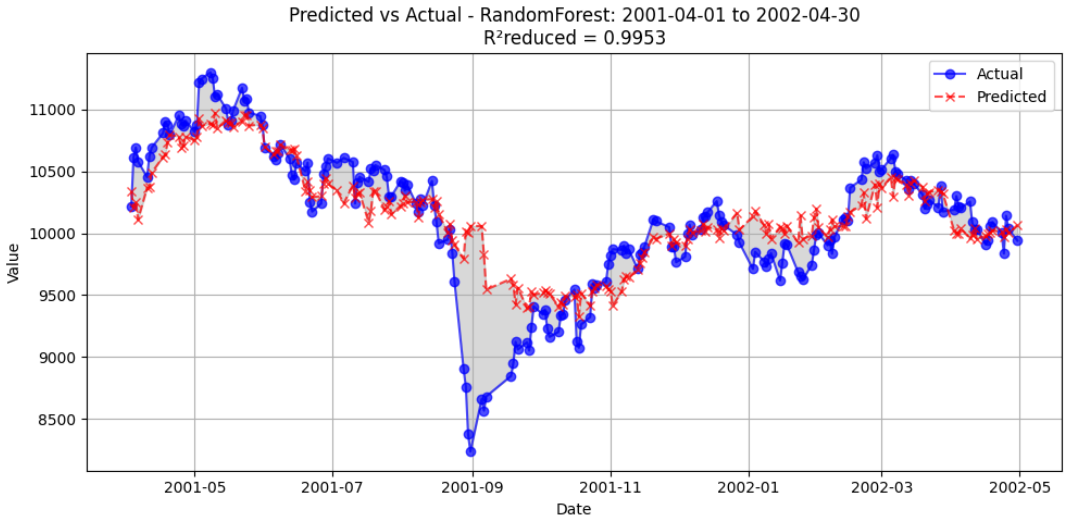


Figura 47: Random forest (32 variabili) – Attacco alle Torri Gemelle (11 settembre)

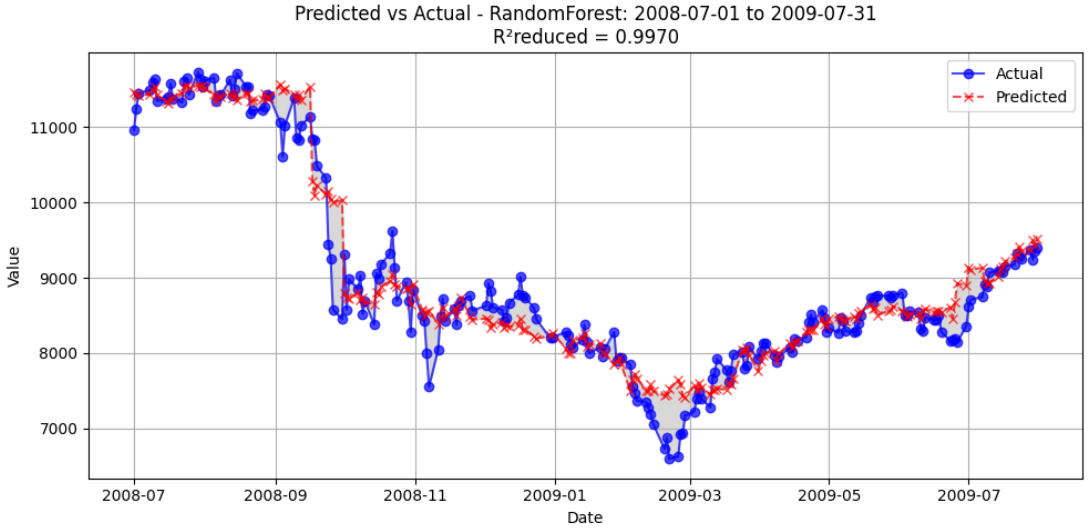


Figura 48: Random forest (32 variabili) – Crisi Lehmann Brothers

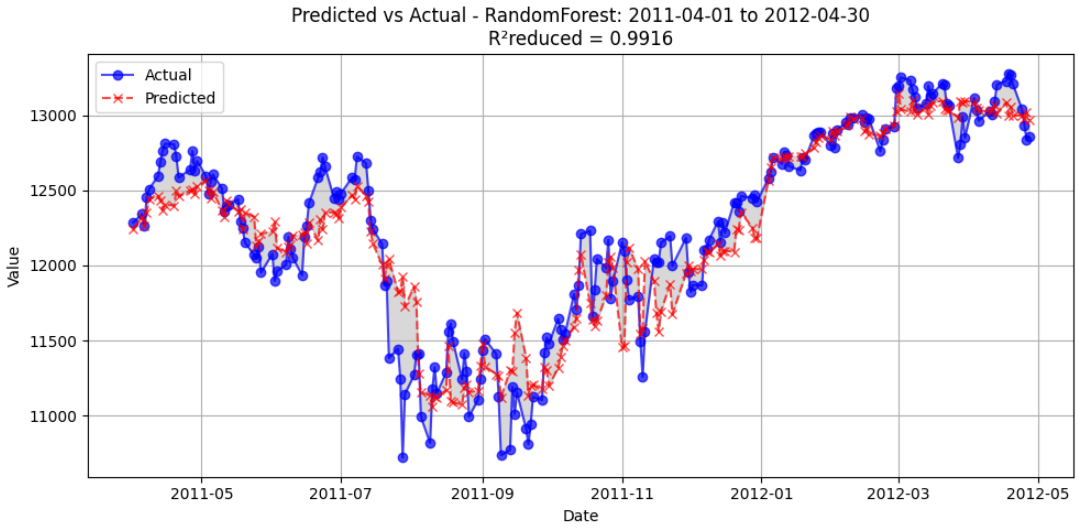


Figura 49: Random forest (32 variabili) – Crisi del debito sovrano

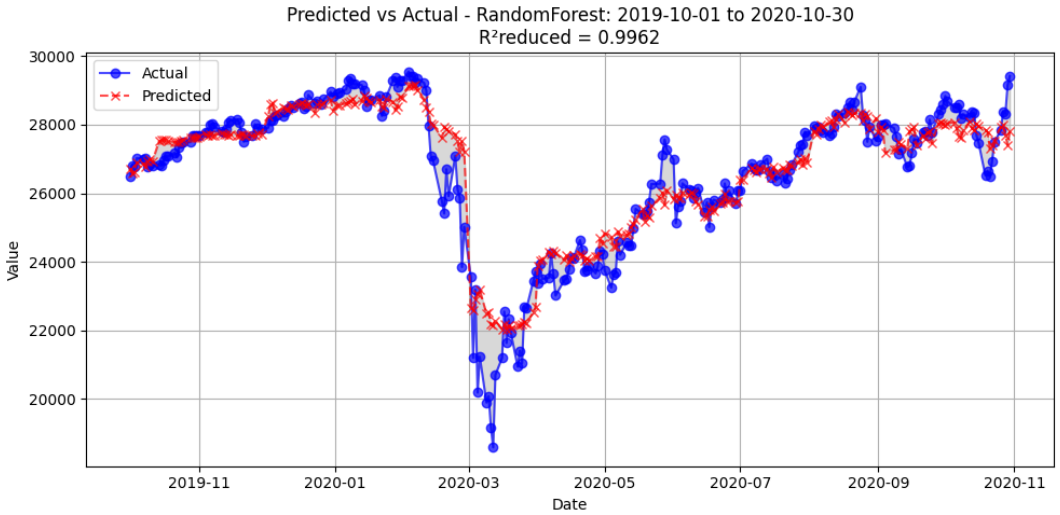


Figura 50: Random forest (32 variabili) – Pandemia COVID-19

7. Conclusioni finali

Lo studio condotto ha dimostrato l'efficacia dell'applicazione di modelli di *machine learning*, supportati da tecniche di ottimizzazione, nella previsione e interpretazione dei movimenti aggregati dei prezzi azionari nei mercati finanziari. Attraverso l'analisi di un ampio set di dati storici, comprendente variabili economiche e finanziarie, è stato possibile identificare le relazioni più significative tra gli indicatori di mercato e l'indice DJIA, consentendo di sviluppare modelli predittivi con elevata accuratezza.

I risultati ottenuti evidenziano come i modelli di regressione lineare multivariata, opportunamente ottimizzati mediante tecniche di selezione delle variabili, abbiano fornito un buon compromesso tra interpretabilità e precisione predittiva. In particolare, l'ottimizzazione basata sull'analisi di correlazione ed autocorrelazione tra le variabili indipendenti, condotta con l'ausilio del software CPLEX di IBM, ha consentito di individuare un set minimo di variabili indipendenti con un'alta capacità esplicativa, senza sacrificare in modo eccessivo le prestazioni del modello. Questi risultati offrono una comprensione più chiara delle interazioni tra variabili economiche e finanziarie, fornendo un valido strumento per l'interpretazione delle dinamiche di mercato. Tuttavia, l'analisi dettagliata ha rivelato alcune limitazioni nella capacità di questi modelli di catturare dinamiche di mercato fortemente non lineari, specialmente in presenza di eventi economici eccezionali.

L'integrazione di un modello di regressione basato su *Random Forest* ha mostrato un netto miglioramento delle prestazioni predittive, con un significativo aumento del coefficiente di determinazione (R^2) e una riduzione dell'errore quadratico medio (MSE). In particolare, questo modello si è rivelato più robusto nell'adattarsi a scenari di mercato caratterizzati da forte volatilità e discontinuità, evidenziando la capacità di gestire in modo più efficace la complessità dei dati finanziari.

Dall'analisi dei risultati si evince che i fattori macroeconomici e finanziari di maggiore rilevanza per la previsione dell'indice DJIA includono la massa monetaria (M2), soprattutto, e poi il tasso di disoccupazione, le variazioni giornaliere dell'indice e le misure di volatilità (VIX e sue derivate). Questi parametri forniscono una base solida per interpretare le dinamiche dei mercati finanziari, interpretando sia le aspettative degli investitori che le condizioni macroeconomiche di fondo. Entrambi giocano un ruolo cruciale nella formazione dei prezzi azionari.

Nonostante i risultati incoraggianti, vi è senz'altro margine per ulteriori sviluppi. In particolare, la capacità dei modelli di prevedere con precisione le crisi finanziarie rimane limitata dalla natura altamente non lineare e imprevedibile di tali eventi. A tal proposito, l'efficacia dei modelli può essere migliorata attraverso l'integrazione di ulteriori fonti di dati, al momento non considerati, come la *sentiment analysis* basata su *news* e *social media*, e l'adozione di approcci di *deep learning* in grado di cogliere *pattern* più complessi.

In conclusione, il presente studio fornisce un *framework* solido e scalabile per l'analisi predittiva dei mercati finanziari, offrendo spunti concreti per future ricerche orientate all'ottimizzazione e all'integrazione di modelli predittivi più avanzati e potenti.

Bibliografia

1. Ando, A. (1963). *An Empirical Model of United States Economic Growth: An Exploratory Study in Applied Capital Theory*. In Models of Income Determination. National Bureau of Economic Research.
2. Bernanke, B. S. (2009). *The Crisis and the Policy Response*. Discorso tenuto alla London School of Economics, 13 gennaio 2009. Disponibile su: https://www.federalreserve.gov
3. Bernanke, B. S. (2020). *The New Tools of Monetary Policy*. American Economic Review, 110(4), 943–983.
   Nota: Approfondimento sulle politiche di Quantitative Easing e le nuove strategie di politica monetaria.
4. Board of Governors of the Federal Reserve System (US). *Market Yield on U.S. Treasury Securities*. [Dati economici]. Disponibile su: https://www.federalreserve.gov
5. Brayton, F., Levin, A., Tryon, R., & Williams, J. C. (1997). *The Role of Expectations in the FRB/US Macroeconomic Model*. Federal Reserve Board, Washington, D.C.
   Nota: Discute il ruolo delle aspettative nei modelli macroeconomici classici e moderni.
6. Brayton, F., & Tinsley, P. A. (1996). *Tying the current price of an asset to its expected future earnings*. In *A Guide to FRB/US: A Macroeconomic Model of the United States*. Federal Reserve Board, Washington, D.C.
7. Chicago Board Options Exchange. *CBOE Volatility Index (VIX)*. [Dati economici]. Disponibile su: https://www.cboe.com
8. Friedman, M., & Schwartz, A. J. (1963). *A Monetary History of the United States, 1867–1960*. Princeton University Press.
   Nota: Opera fondamentale per comprendere il rapporto tra offerta di moneta e livelli dei prezzi.
9. Friedman, M. (1969). *The Optimum Quantity of Money*. Aldine Transaction.
   Nota: Contiene la provocatoria teoria della "*Helicopter Money*".
10. Keynes, J. M. (1936). *The General Theory of Employment, Interest, and Money*. London: Macmillan.
    Nota: Fondamentale per le teorie keynesiane sull'intervento statale nell'economia e sulla trappola della liquidità.
11. Modigliani, F., & Ando, A. (1963). *The "Life Cycle" Hypothesis of Saving: Aggregate Implications and Tests*. American Economic Review, 53(1), 55-84.
12. Roosevelt, F. D. (1933-1939). *The New Deal*. Serie di programmi economici e sociali implementati durante la Presidenza Roosevelt per affrontare la Grande Depressione. Per approfondimenti storici si veda:
    - Kennedy, D. M. (1999). *Freedom from Fear: The American People in Depression and War, 1929-1945*. Oxford University Press.
    - Schlesinger Jr., A. M. (1957). *The Age of Roosevelt*. Boston: Houghton Mifflin.
13. U.S. Bureau of Labor Statistics. *Consumer Price Index for All Urban Consumers (CPI-U)*. [Dati economici]. Disponibile su: https://www.bls.gov
14. U.S. Bureau of Economic Analysis. *Real Gross Domestic Product*. [Dati economici]. Disponibile su: https://www.bea.gov

15. U.S. Department of the Treasury. *Federal Debt: Total Public Debt*. [Dati economici]. Disponibile su: https://www.fiscal.treasury.gov

16. Federal Reserve Bank of St. Louis. *St. Louis Fed Financial Stress Index*. [Dati economici]. Disponibile su: https://fred.stlouisfed.org

17. *Macrotrends*. S&P 500 P/E Ratio. [Dati economici]. Disponibile su: https://www.macrotrends.net

18. Skidelsky, R. (2009). *Keynes: The Return of the Master*. London: Penguin Books.

19. Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

20. Visco, I. (1984). *Dalla Teoria Alla Pratica Nei Modelli Macroeconomici: L'Eclettismo Post-Keynesiano*. Moneta e Credito, 37(147), 347-371.

21. Zou, H., & Hastie, T. (2005). *Regularization and Variable Selection via the Elastic Net*. Journal of the Royal Statistical Society: Series B, 67(2), 301-320.

22. *Investing.com*. Dow Jones Industrial Average (DJIA) Historical Data. [Dati di mercato]. Disponibile su: https://www.investing.com

23. *Ritter, J. R. (2005). Economic Growth and Equity Returns. Pacific-Basin Finance Journal, 13(5), 489-503.*
Nota: Analizza la relazione tra aspettative di crescita e multipli di valutazione, incluso il P/E.

24. Kumar, I., Dogra, K., Utreja, C., & Yadav, P., 2018. *A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1003-1007. Disponibile su: https://doi.org/10.1109/ICICCT.2018.8473214.

25. Tabachnick BG, Fidell LS (2001) *Using Multivariate Statistics 4th ed*. (Allyn and Bacon, Boston).

26. Massy WF (1965) *Principal components regression in exploratory statistical research*. J. Amer. Statist. Assoc. 60(309):234–256.

27. Wold S, Ruhe A, Wold H, Dunn W III (1984) *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Scientific Statist. Comput.* 5(3):735–743.

Glossario

CBOE ***C**hicago **B**oard **O**ptions **E**xchange*

C_{max} *Valore massimo del coefficiente di correlazione a coppie*. Parametro che definisce la soglia massima di correlazione accettabile tra due variabili indipendenti in un modello di regressione. Se due variabili presentano una correlazione superiore C_{max}, una delle due viene rimossa per ridurre la multicollinearità, migliorando così la stabilità e l'interpretabilità del modello.

CPI ***C**onsumer **P**rice **I**ndex*. Il **CPI**, ossia l'indice dei prezzi al consumo, è una media dei prezzi di un opportuno paniere di beni di consumo e servizi che vengono utilizzati per il calcolo dell'inflazione.

DJIA ***D**ow **J**ones **I**ndustrial **A**verage*: è un indice composto dai 30 titoli principali del mercato azionario statunitense pesati sulla base del loro prezzo. È un indice "*price weighted*". I settori più importanti sono i seguenti:

- *information technology:* 26,30%
- salute: 14,50%;
- consumi discrezionali: 14,20%;
- industria: 14%;
- finanza: 13,30%.

ESS ***E**xplained **S**um of **S**quares*: misura della devianza spiegata dal modello

FED La ***FED**eral Reserve Bank* (o ***FED***), è la banca centrale responsabile della stabilità monetaria e finanziaria negli Stati Uniti. Fa parte di un sistema più ampio, noto come *Federal Reserve System*, con 12 banche centrali regionali, situate nelle principali città degli Stati Uniti.

FRB Il modello econometrico **FRB** (*Federal Reserve Board*) è un modello macroeconomico sviluppato dalla *Federal Reserve* per analizzare l'economia statunitense, monitorare l'inflazione e supportare le decisioni di politica monetaria. A differenza del modello MPS, il FRB è stato progettato specificamente per riflettere le priorità e i meccanismi della politica monetaria, con una maggiore enfasi sulle dinamiche dei tassi di interesse, della domanda aggregata e dei mercati finanziari. Le differenze principali rispetto al modello MPS sono:

1. **Focus istituzionale**: Il FRB è orientato alle esigenze della Federal Reserve, con un'attenzione particolare alle dinamiche monetarie e al sistema bancario. Il MPS, pur considerando la politica monetaria, ha un approccio più generale ed è utilizzato anche per valutare gli effetti di politiche fiscali.

2. **Evoluzione e aggiornamenti**: Il FRB è stato continuamente aggiornato per includere nuove metodologie e incorporare i cambiamenti strutturali dell'economia, come la globalizzazione e l'innovazione finanziaria. Il modello MPS, invece, è stato pressoché abbandonato negli anni '80, essendo superato dai progressi nei modelli econometrici e computazionali.

3. **Struttura tecnica**: Il modello FRB si basa su approcci più moderni, inclusi i modelli DSGE (*Dynamic Stochastic General Equilibrium*), che combinano teorie microeconomiche e macroeconomiche per rappresentare il comportamento degli agenti economici in modo più realistico. Il MPS utilizza un sistema di equazioni simultanee più tradizionale e statico.

GDP ***G**ross **D**omestic **P**roduct*: prodotto interno lordo.

M2 **M2** è una misura della massa monetaria che rappresenta l'insieme di tutto il denaro in circolazione in un'economia, includendo componenti con diversi livelli di liquidità. Comprende:

- moneta fisica in circolazione (banconote e monete);
- depositi a vista nei conti correnti bancari.
- altri strumenti di pagamento immediato (ad es. assegni);
- depositi a risparmio: fondi che possono essere prelevati su richiesta, ma che spesso non sono utilizzati per i pagamenti giornalieri;
- depositi a breve termine (certificati di deposito e altri strumenti di risparmio vincolati per brevi periodi).

M2 è spesso usato per analizzare la liquidità complessiva dell'economia e per prevedere inflazione, tassi di interesse e crescita economica.

MPS Il modello econometrico **MPS** (***M****assachusetts Institute of Technology, University of* ***P****ennsylvania,* ***S****ocial Science Research Council,* dai nomi degli istituti universitari ove operavano i suoi principali sviluppatori, ossia, rispettivamente, Franco Modigliani e Albert Ando, e dall'organizzazione tramite la quale il modello è stato aggiornato e gestito per conto della *Federal Reserve*) è un modello macroeconomico sviluppato negli anni '60 per analizzare e prevedere il comportamento dell'economia statunitense. È uno dei primi modelli econometrici di grandi dimensioni, basato su equazioni simultanee che descrivono le relazioni tra variabili economiche chiave come consumo, investimenti, occupazione, inflazione e commercio estero.

Il modello MPS è stato utilizzato per simulazioni e analisi delle politiche economiche, consentendo di studiare gli effetti di cambiamenti nelle politiche fiscali e monetarie sull'economia complessiva. Sebbene oggi sia considerato superato rispetto ai modelli econometrici moderni, rappresenta un importante passo nello sviluppo dell'analisi quantitativa delle economie nazionali.

MAE ***M****ean* ***A****bsolute* ***E****rror*

MSE ***M****ean* ***S****quare* ***E****rror*

NaN ***N****ot-****A****-****N****umber*: identificativo convenzionale per dati il cui tipo non è interpretabile in modo numerico

R^2 Coefficiente di Determinazione

R^2 *adjusted* Coefficiente di Determinazione Adattato

RMSE ***R****oot* ***M****ean* ***S****quared* ***E****rror*

RSS ***R****esidual* ***S****um of* ***S****quares*: misura della devianza dei residui

TSS ***T****otal* ***S****um of* ***S****quares*: misura della devianza totale

S&P500 **S**tandard**&P**oor**500**: è un indice composto da 500 azioni a larga capitalizzazione e rappresenta circa l'80% dell'intero mercato azionario statunitense. È un indice "*value weighted*", ossia il peso delle società che lo compongono cambia in base alla capitalizzazione delle stesse. Di seguito, in sintesi, i settori più importanti dell'indice:

- *information technology*: 27,50%;
- salute: 14,60%;

- consumi discrezionali: 10,80%;
- servizi di comunicazione: 10,80%;
- finanza: 10,10%;
- industria: 8%.

SVR ***S**upport **V**ector **R**egression* è una tecnica di *machine learning* basata sul metodo delle macchine a vettori di supporto (*Support Vector Machines*, SVM), progettata per problemi di regressione anziché di classificazione. SVR si distingue per la capacità di trovare una funzione di previsione che bilancia l'accuratezza e la complessità, ottimizzando una tolleranza agli errori definita dall'utente.

VIX **V**olatility **I**nde**X**: indice in tempo reale che rappresenta le aspettative del mercato sulla forza relativa delle variazioni di prezzo a breve termine dell'indice S&P 500. Si ricava a partire dai prezzi delle opzioni sull'S&P 500 con scadenza a breve termine.