Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/

https://doi.org/10.26219/heal.aueb.6444

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Master of Science in Statistics

Modeling oil market series using econometric models and machine learning techniques

Anastasios Bekas

Supervisor: **Ioannis Vrontos**

Athens

November 2024

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Acknowledgements

I would like to extend my heartfelt gratitude to my dear parents for their unwavering emotional and material support. Additionally, I wish to express my deepest appreciation to my supervising professor, I. Vrontos, for the countless hours dedicated to enhancing my understanding of mathematical and theoretical concepts, and for the enthusiasm that motivated me to pursue my research in this field with great passion.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Abstract

This thesis examines the dynamics of crude oil prices, focusing on the returns of WTI, Brent, and Dubai benchmarks. By integrating econometric models such as ARMA-GARCH with machine learning techniques like random forests and support vector regression, the study captures the complexities of oil price behavior, including non-linear dependencies and volatility clustering. The analysis incorporates structural breaks from significant global events, such as the financial crisis and the COVID-19 pandemic, to enhance forecasting accuracy. External macroeconomic factors and policy uncertainty indices are identified as key determinants influencing oil market dynamics. The findings highlight the value of combining traditional and modern methodologies to address the intricacies of energy markets. The study concludes with practical insights for policymakers and market participants, emphasizing the importance of hybrid approaches and the potential of alternative data sources in improving forecasting models.

Keywords: Crude oil prices, WTI, Brent, Dubai, econometric models, machine learning, ARMA-GARCH, random forest, support vector regression, structural breaks, volatility clustering, macroeconomic indicators, policy uncertainty, forecasting.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τη δυναμική των τιμών του αργού πετρελαίου, εστιάζοντας στις αποδόσεις των δεικτών WTI, Brent και Dubai. Μέσω της ενσωμάτωσης οικονομετρικών μοντέλων, όπως ARMA-GARCH, με τεχνικές μηχανικής μάθησης, όπως Random forest, Support Vector Regression, η μελέτη αναλύει την πολυπλοκότητα της συμπεριφοράς των τιμών του πετρελαίου, περιλαμβάνοντας μη γραμμικές εξαρτήσεις και συστάδες μεταβλητότητας. Η ανάλυση λαμβάνει υπόψη τις δομικές αλλαγές από σημαντικά παγκόσμια γεγονότα, όπως η χρηματοπιστωτική κρίση και η πανδημία COVID-19, για τη βελτίωση της ακρίβειας των προβλέψεων. Εξωτερικοί μακροοικονομικοί παράγοντες και δείκτες αβεβαιότητας πολιτικής εντοπίζονται ως καθοριστικοί παράγοντες που επηρεάζουν τη δυναμική της αγοράς πετρελαίου. Τα ευρήματα υπογραμμίζουν την αξία του συνδυασμού παραδοσιακών και σύγχρονων μεθοδολογιών για την αντιμετώπιση των σύνθετων προκλήσεων των αγορών ενέργειας. Η μελέτη καταλήγει σε πρακτικά συμπεράσματα για τους φορείς χάραξης πολιτικής και τους συμμετέχοντες στην αγορά, τονίζοντας τη σημασία των υβριδικών προσεγγίσεων και τις δυνατότητες των εναλλακτικών πηγών δεδομένων για τη βελτίωση των μοντέλων πρόβλεψης.

Λέξεις-Κλειδιά: Τιμές αργού πετρελαίου, WTI, Brent, Dubai, οικονομετρικά μοντέλα, μηχανική μάθηση, ARMA-GARCH, Random Forest, Support Vector Regression, διαρθρωτικές αλλαγές, συστάδες μεταβλητότητας, μακροοικονομικοί δείκτες, αβεβαιότητα πολιτικής, πρόβλεψη.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Contents

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

List of Figures

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

List of Tables

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 1

Introduction

Energy is necessary for modern life since it allows for the heating of buildings and the operation of factories. The industrialization of contemporary economies around the world has been greatly aided by the discovery and widespread usage of nonrenewable energy sources including coal, crude oil, and natural gas. Even though the use of renewable energy has increased recently, conventional energy sources still play a vital role in the economy. The circumstances surrounding the world's energy markets have been unparalleled in the last year. Due to a lack of supply and excessive demand brought on by the COVID-19 pandemic and, more recently, Russia's full-scale invasion of Ukraine, energy commodity prices have reached all-time highs. Global economies have been greatly impacted by these events. The soaring prices of natural gas, crude oil, and electricity have profoundly affected inflation, economic growth, living standards, and broader policy objectives like decarbonization. These price hikes have created significant challenges for economies trying to balance growth with sustainability and affordability (OECD, 2022).

Understanding the behavior of volatility in crude oil prices is crucial for several reasons. Fluctuations in volatility can significantly impact the risk exposure of both oil producers and industrial consumers, potentially influencing their investments in oil production assets, reserves, inventories, and related facilities. Additionally, volatility plays a key role in determining the value of commodity-based contingent claims. Consequently, a deep understanding of volatility dynamics is essential for making informed decisions about derivative valuation, hedging strategies, and investments in the oil sector. Moreover, there is substantial evidence that volatility in oil prices can have a ripple effect on broader financial markets and the overall economy. Therefore, both public and private sector policymakers, as well as participants in financial markets, can greatly benefit from understanding how unexpected events—whether positive or negative—impact oil price volatility (Ewing and Malik, 2017).

Crude oil prices have experienced significant volatility in recent years (see Fig. 1.1). For instance, the price of WTI oil soared to 133.88 per barrel in June 2008, only to plunge dramatically to 41.12 per barrel by December of the same

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

year, largely due to the financial crisis. Additionally, the COVID-19 pandemic caused an unprecedented negative demand shock, resulting in oil prices dropping to negative levels in April 2020, with prices hitting as low as 16.55 per barrel (Liu et al., 2022).

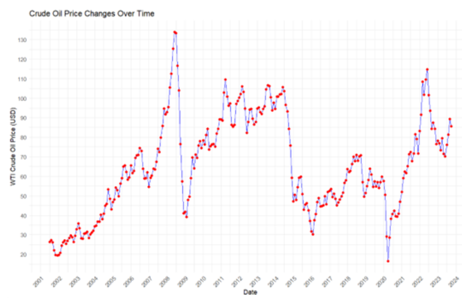

Figure 1.1: Crude oil monthly prices (WTI).

While changes in oil prices are widely recognized for their significant effects, recent research has increasingly focused on the implications of oil price volatility. Oil price volatility is a critical factor in macroeconomic models and plays a substantial role in determining the prices of oil futures. For instance, Ferderer (1996) provides empirical evidence demonstrating that oil price volatility significantly influences aggregate output fluctuations in the United States. Similarly, Sadorsky (1999) finds that both oil prices and their volatility are crucial in affecting stock returns. Guo and Kliesen (2005) offer robust evidence that oil price volatility has a notably negative impact on U.S. GDP growth. Additionally, in a more recent study, Malik and Ewing (2009) identify significant transmission of shocks and volatility between oil prices and equity sector returns, attributing these results to the concept of cross-market hedging.

Given the complexities of the oil market and the critical role of oil in the global economy, accurate forecasting of oil prices and volatility has become increasingly important. Traditional econometric models, such as Generalized Autoregressive Conditional Heteroskedasticity (GARCH) developed by Bollerslev (1986), have been extensively used to model and predict the volatility of financial time series, including crude oil prices. These models help capture the persistence and clustering of volatility, which are common features in financial data. However, the dynamic and often non-linear nature of oil markets has led to the exploration of more sophisticated techniques.

In recent years, machine learning (ML) approaches have gained traction as powerful tools for forecasting in complex environments like energy markets. Techniques such as random forests, support vector machine, gradient boosting, and deep learning models, including long short-term memory (LSTM) networks and temporal fusion transformers, have shown promise in capturing intricate patterns in time series data that traditional models might miss. These ML models, as highlighted by Zhang et al. (2017), can automatically learn from vast amounts of data, adjusting for non-linearity and interactions between variables.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

thereby improving the accuracy of predictions.

The integration of machine learning techniques with traditional econometric methods has opened new avenues for research and application in energy market forecasting. By leveraging the strengths of both approaches, researchers and practitioners can develop more robust models that provide deeper insights into market dynamics. This is particularly important as the energy market becomes increasingly influenced by external factors such as geopolitical events, technological advancements, and policy shifts aimed at transitioning towards sustainable energy sources.

As energy markets continue to evolve, understanding and predicting oil price volatility remains a critical challenge for economists, investors, and policymakers. Accurate forecasts are essential not only for managing risks associated with price fluctuations but also for making informed decisions regarding investment in energy infrastructure, developing hedging strategies, and formulating policies that ensure economic stability and growth. The ongoing advancements in econometric modeling and machine learning offer exciting opportunities to enhance our understanding of oil market dynamics and improve the precision of forecasts in this vital sector.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 2

Literature review

2.1 Problem Statement

Forecasting oil prices and their volatility has long been a challenge of critical importance for policymakers, investors, and industries reliant on energy markets. The complexity of oil price dynamics arises from their sensitivity to global economic trends, geopolitical developments, and financial market conditions. In recent years, the increasing availability of high-frequency data and the evolution of econometric and machine learning methodologies have opened new pathways for tackling this challenge.

The study supplements traditional econometric techniques and machine learning methods to assess the oil market. We illustrate this by applying the Bai-Perron test of structural breaks to the oil returns series, which known to be driven by major economic and geopolitical events, such as that in 2008 (the global financial crisis) and in 2020 (Covid-19 pandemic shocks to oil prices). Identifying those segments allows us to model oil price behaviour more accurately under different conditions, as introduced by Bai and Perron (1998).

Traditional econometric methods, such as stepwise regression and models with autocorrelated errors, are employed to capture linear relationships and ensure model parsimony. For example, the selection of statistical significant variables, reflects the importance of global economic uncertainty and regional policy measures in influencing oil prices. Furthermore, non-linear modeling techniques are used to explore asymmetric dependencies and tail risk behavior, inspired by works such as Koenker and Hallock (2001).

Advanced machine learning techniques are also integrated into the analysis. Inspired by state-of-the-art methodologies, this research employs Random Forest, Support Vector Regression (SVR), and Gradient Boosting to enhance predictive accuracy. These models are well-suited for capturing complex relationships and non-linear dependencies in the oil market data, providing robust insights into its temporal dynamics Breiman (2001), Smola and Schölkopf (2004), Friedman (2001). These methods are complemented by traditional GARCH-

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

family models (Engle, 1982; Bollerslev, 1986) to account for heteroskedasticity in returns, capturing volatility clustering often observed in oil prices.

A unique contribution of this thesis is its focus on quantifying the impact of significant geopolitical events, including the Ukraine conflict and major economic crises, on crude oil returns. By employing a comprehensive dataset that includes global crude oil prices (Brent, WTI, and Dubai) as dependent variables and various macroeconomic and market factors (such as equity market volatility, economic policy uncertainty indices for Europe and the United States, general business conditions, and treasury rates) as explanatory variables, this analysis offers a multifaceted perspective.

Incorporating these diverse predictors, including the Federal Funds Effective Rate, Treasury Bill rates, and global price of Nickel, the study explores the complex interplay between global uncertainty and energy markets. This approach provides valuable insights with practical implications for risk management, investment strategies, and policy formulation in the face of heightened uncertainty.

Through the integration of econometric and machine learning techniques, this research aims to deepen our understanding of oil price dynamics and improve predictive accuracy in the face of global uncertainties. By addressing the methodological gaps in existing literature and introducing robust forecasting tools, this study contributes to the evolving field of energy economics and financial econometrics.

2.2 Importance of Study

The ability to accurately forecast oil price movements and their volatility is crucial, ranging from policymakers and investors to industry participants. The fluctuations in oil prices have significant implications for global economic stability, energy security, and the functioning of financial markets. Effective predictive models provide essential insights that support decision-making, risk management, and strategic planning (Li and Ge, 2013; Lahmiri, 2017).

The oil market is influenced by a complex interplay of factors, including geopolitical tensions, economic cycles, and technological advancements. This complexity necessitates the development of sophisticated analytical tools capable of capturing the non-linear and non-stationary characteristics of oil prices. Traditional econometric models, while valuable, often fall short in fully capturing these intricate relationships and the sudden, often unexpected shifts that occur in today's energy markets (Yu et al., 2016; Ewing and Malik, 2017).

This research is particularly important as it explores both conventional econometric methods and modern machine learning techniques to enhance the accuracy of oil price predictions. By assessing the effectiveness of these approaches in the context of recent global events, such as the Ukraine conflict, this study aims to deepen our understanding of the driving forces behind oil price volatility. The results of this research have the potential to improve forecasting models, thereby enhancing market efficiency, reducing uncertainty, and

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

contributing to the development of more resilient energy policies (Liu et al., 2022).

2.3 Review of Relevant Research

The complex nature of oil price movements has consistently drawn the interest of economists and policymakers because of the vital role this commodity plays in global markets and economic stability. Over the years, a vast body of literature has developed, employing an array of methodologies to address the challenges of forecasting oil prices in an ever-evolving market environment. These approaches have progressed from traditional econometric models, which rely on linear assumptions, to advanced machine learning techniques capable of capturing complex and non-linear dynamics. This section provides a critical review of these methodologies, highlighting their strengths and limitations, and examines how external shocks, such as geopolitical events, influence oil price volatility and structural market dynamics.

This review also situates the present study within the broader context of forecasting research, illustrating how the integration of econometric and machine learning models offers a novel perspective on understanding and predicting oil price movements.

2.3.1 Traditional Econometric Approaches

Traditional econometric models, such as autoregressive integrated moving average (ARIMA) models, which are intended to identify trends in economic data across time, were mostly employed in the early studies on oil price forecasting. These models, which are based on time series analysis and linear regression, have been used extensively because they are easy to use and good at simulating oil price trends and seasonality. However, ARIMA is less successful at capturing the volatility and abrupt shifts frequently observed in oil markets since it makes the assumption that oil prices follow a linear path.

In an attempt to increase predicting accuracy, researchers like Li and Ge (2013) have expanded standard regression frameworks to incorporate extra predictors including sentiment measures, macroeconomic indicators, and geopolitical events. These developments have improved our knowledge of the variables affecting oil pricing. However, the linear nature of these models remains a limitation, as they cannot adequately capture the non-linear and unpredictable changes in oil price fluctuations.

In our analysis, traditional econometric models are used as a baseline for comparison with more advanced methods. While these models offer useful insights into long-term trends and relationships, their inability to adapt to sudden market changes and high-frequency data highlights the need for more flexible approaches. This has motivated the use of machine learning and non-linear techniques, which are discussed in later sections of this work.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

2.3.2 Non-linear Models and Hybrid Approaches

As the limitations of purely linear models in capturing the complexities of oil price movements became evident, researchers have increasingly turned to non-linear models. Among these, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, introduced by Engle (1982), have been widely adopted. GARCH models are particularly suited for handling volatility clustering, a characteristic feature of financial time series where periods of high volatility are followed by periods of low volatility. Multivariate extensions of GARCH, proposed by Bollerslev (1986), further allow for the modeling of correlations between multiple variables, enhancing their applicability in interconnected markets like crude oil.

By combining linear and non-linear techniques, hybrid approaches have proven to be effective at improving forecasting accuracy alongside GARCH models. For example, Yu et al. (2016) combined traditional econometric models with a machine learning method called least squares support vector regression (LSSVR). These hybrid frameworks leverage the strengths of both methods, maintaining the interpretability of linear models while using machine learning to uncover non-linear patterns.

To improve predictive performance, our analysis builds on this by utilizing a variety of machine learning approaches, such as gradient boosting, support vector regression (SVR), and random forests. These techniques excel at managing complicated interactions and non-linear connections between predictors. This study provides a comprehensive structure for understanding and predicting the dynamics of oil prices by combining such machine learning approaches with econometric models, taking into account both slow trends and sudden market fluctuations.

2.3.3 Machine Learning Techniques

The rapid development of big data and computational advancements has facilitated the application of machine learning techniques to oil price forecasting. These methods, which include decision trees, support vector regression (SVR), and ensemble learning approaches, offer significant advantages over traditional models by being capable of capturing complex, non-linear relationships in large and dynamic datasets.

In our analysis, decision tree-based methods such as random forests and gradient boosting machines have been applied to forecast oil prices. These models are particularly effective at handling large datasets with numerous predictors, allowing for the analysis of the wide array of factors influencing oil prices. Their ability to rank the importance of predictors also provides valuable insights into the key drivers of oil market dynamics.

Support vector regression (SVR) has also been employed, offering robustness in scenarios where data may exhibit non-linear trends and high variability. This technique's foundation in statistical learning theory enables it to perform well in forecasting tasks by balancing model complexity and predictive accuracy.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Furthermore, research by Liu et al. (2022) demonstrates the integration of quantile regression with ensemble empirical mode decomposition (EEMD) to forecast oil prices. Such approaches are particularly useful for exploring the tails of the price distribution, enabling a more detailed understanding of oil price movements and volatility under varying market conditions. By combining these advanced machine learning techniques with structural break analysis, our research provides a comprehensive framework for forecasting oil prices in the face of both gradual market trends and sudden shocks.

2.3.4 Impact of External Shocks and Geopolitical Events

A critical aspect of oil price forecasting is the ability to account for external shocks, such as geopolitical events, which often cause significant fluctuations in oil prices. The literature has extensively documented the effects of events such as wars, natural disasters, and economic crises on energy markets. For example, the recent conflict in Ukraine and related geopolitical tensions have profoundly impacted global oil markets, influencing both prices and volatility.

Research by Ewing and Malik (2017) emphasizes the importance of incorporating external shocks into forecasting models, particularly in the context of oil and natural gas markets, which exhibit interconnected dynamics. These studies suggest that traditional econometric models may struggle to capture the rapid and often asymmetric responses of oil prices to external events, highlighting the need for more flexible and adaptive forecasting approaches.

Machine learning methods like support vector regression (SVR) and random forest have been used in our investigation to evaluate the short-term and long-term effects of external shocks. More accurate predictions can be made during times of increased volatility thanks to these methodologies' greater adaptability to real-time data changes. Through the integration of high-frequency data and structural break analysis, these models offer a more sophisticated comprehension of the ways in which geopolitical events impact the dynamics of oil prices. For market participants trying to manage the uncertainty of external shocks, this strategy provides a useful tool.

2.3.5 Comparative Studies and Model Performance Evaluation

The comparison of modern machine learning methods to traditional econometric models has been a major area of study in oil price predictions. These comparative studies are essential for identifying specific situations in which each methodology works best, which helps to direct future developments in the area. For example, Yu et al. (2016) performed a thorough investigation contrasting the effectiveness of conventional ARIMA models with least squares support vector regression (LSSVR). According to their research, LSSVR continuously performed better than ARIMA in terms of accuracy and robustness, particularly when it came to identifying the non-linear correlations present in oil price data. Similarly, Lahmiri (2017) compared artificial neural networks (ANNs)

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

and support vector machines (SVMs), finding that ANNs were able to model complicated dependencies in the data, which led to more accurate forecasts.

To evaluate model performance, these studies usually use evaluation metrics including mean squared error (MSE), mean absolute error (MAE), and R-squared values. Our investigation compares models like support vector regression (SVR), random forests, and gradient boosting using comparable criteria, showing that applying machine learning techniques to structural break-adjusted datasets significantly improves prediction accuracy. In addition to providing useful advice for investors making data-driven decisions, this set of research increases our understanding of oil price dynamics by carefully comparing models across various time periods and datasets.

2.3.6 Future Directions and Emerging Trends

Oil price forecasting is a field that is always changing to meet the demands of a world economy that is becoming more connected and complex. The creation of hybrid systems by combining econometric models and artificial intelligence (AI) is one significant development. These methods combine the interpretability of traditional econometric frameworks with the predictive capability of machine learning algorithms like gradient boosting and support vector regression (SVR). The processing of both structured and unstructured data, such as news mood and geopolitical signals, which provide deeper insights into market dynamics and possible price movements, is where these hybrid models excel.

Last but not least, the interaction between established oil markets and developing alternative energy markets is becoming a crucial area of attention as the world's energy industry experiences an important transition towards renewable sources. Although they are still in their beginnings, forecasting models that take into account factors from sustainability indices, carbon pricing, and increases in renewable energy are an important field for further study. These models will be crucial for directing investment plans, forming energy policy, and guaranteeing a more seamless shift to a diversified energy future.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 3

Econometric Modeling Approaches

3.1 Autoregressive Model (AR)

The Autoregressive (AR) model is a fundamental time series model that explains the current value of a variable based on its own past values. An AR model of order 1, denoted as AR(1), can be expressed as:

$$y_t = \alpha_0 + \alpha_1 \cdot y_{t-1} + \epsilon_t \tag{3.1}$$

Y_t represents the value of the series at time t, α_0 is the intercept, and α_1 is the coefficient capturing the influence of the previous value y_{t-1} on the current observation. The term ϵ_t is white noise, which is assumed to have a mean of zero and a constant variance σ^2.

For a higher-order AR process, denoted as AR(p), the model incorporates additional lags and is written as:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \epsilon_t \tag{3.2}$$

P indicates the number of lagged terms included in the model, allowing it to account for more complex temporal dependencies.

The AR model is particularly useful for capturing persistence in time series data, where current values are influenced by their historical patterns. It is important to note that the AR model assumes stationarity, meaning that the statistical properties of the series—such as mean, variance, and autocorrelation—remain constant over time. Before applying the AR model, stationarity is typically verified using tests like the Augmented Dickey-Fuller (ADF) test.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

3.2 Moving Average Model (MA)

The Moving Average (MA) model is a time series model that explains the current value of a variable as a function of past error terms. An MA model of order 1, denoted as MA(1), can be expressed as:

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \epsilon_t \tag{3.3}$$

Y_t represents the value of the series at time t, μ is the mean of the series, and θ_1 is the coefficient capturing the influence of the previous error term ϵ_{t-1} on the current value. The term ϵ_t is white noise, which is assumed to have a mean of zero and a constant variance σ^2.

For a higher-order MA process, denoted as MA(q), the model incorporates more past error terms and is written as:

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{3.4}$$

q indicates the number of lagged error terms included in the model. Unlike the Autoregressive (AR) model, the MA model does not directly depend on previous values of y_t, but rather on the residual errors from prior observations.

The MA model is particularly effective for modeling short-term dependencies in time series data and for capturing sudden shocks or noise patterns. It is often combined with AR models to form the ARMA or ARIMA frameworks for more comprehensive time series analysis. As with AR models, stationarity of the data is a key assumption for the application of the MA model.

3.3 Autoregressive Moving Average Model (ARMA)

The Autoregressive Moving Average (ARMA) model combines the features of the Autoregressive (AR) and Moving Average (MA) models, making it a powerful tool for modeling time series data. An ARMA model of order (p, q) incorporates p lagged values of the series and q lagged error terms. The general form of the ARMA(p, q) model is given as:

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{3.5}$$

- y_t: The value of the series at time t.
- α_0: The intercept term.
- α_i: Coefficients for the i-th lagged values of y_t.
- θ_j: Coefficients for the j-th lagged error terms.
- ϵ_t: White noise with a mean of zero and constant variance σ^2.

The ARMA model combines the strengths of AR and MA processes:

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- The **AR component** captures the influence of past values (y_{t-i}) on the current value.
- The **MA component** captures the influence of past error terms (ϵ_{t-j}) on the current value.

The ARMA model is suitable for stationary time series, where the statistical properties such as mean and variance remain constant over time. Before applying an ARMA model, it is essential to test for stationarity, often using the Augmented Dickey-Fuller (ADF) test or other similar methods.

The flexibility of the ARMA model makes it a widely used tool in time series analysis, capable of capturing both long-term dependencies and short-term noise patterns in data.

3.4 Autoregressive Conditional Heteroskedasticity (ARCH)

The Autoregressive Conditional Heteroskedasticity (ARCH) model is a statistical approach developed to analyze and model time series data with volatility that changes over time. This phenomenon, known as heteroskedasticity, occurs when the conditional variance of a series is not constant but instead varies in response to past disturbances. The ARCH model addresses this by allowing the variance of the current error term to depend on the magnitudes of previous periods' error terms.

One of the key features of the ARCH model is its ability to capture volatility clustering, a common characteristic of financial time series. Volatility clustering refers to the tendency of large changes in asset prices to follow other large changes, and small changes to follow other small changes, creating periods of high and low volatility.

The ARCH(q) model specifies the conditional variance σ_t^2 of the error term at time t as a function of the squared error terms from the previous q periods. Mathematically, it is expressed as:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2 \tag{3.6}$$

- σ_t^2: The conditional variance of the error term at time t.
- ϵ_{t-i}^2: The squared error terms from previous time periods $(t-i)$.
- a_0: A constant term ensuring that the variance remains positive.
- a_i: Coefficients that measure the influence of past squared error terms.
- q: The order of the ARCH model, representing the number of lagged squared error terms included.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

By using past squared errors to model current variance, the ARCH model provides a framework to understand how historical volatility impacts current market dynamics. It is particularly useful for financial data where changes in asset prices are often unpredictable but tend to exhibit clustered volatility.

The ARCH model is often used as a foundation for more complex models, such as the Generalized ARCH (GARCH), which further refines the modeling of volatility patterns by including lagged conditional variances in the specification.

3.5 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model was developed to address the limitations of the ARCH model. While the ARCH model effectively captures time-varying volatility, it often requires a large number of parameters to model highly volatile series, leading to inefficiency. The GARCH model extends ARCH by introducing a more compact and flexible structure, making it better suited for analyzing financial time series data with persistent volatility patterns.

The key improvement of the GARCH model lies in its ability to model the conditional variance as a function of both past squared error terms (as in the ARCH model) and past conditional variances. By including lagged variances, the GARCH model captures the persistence of volatility over time with fewer parameters, addressing the inefficiency of the ARCH framework for long-memory processes.

The GARCH(p, q) model is defined as:

$$\sigma_t^2 = a_0 + \sum_{j=1}^{q} a_j \epsilon_{t-j}^2 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2 \tag{3.7}$$

- σ_t^2: The conditional variance of the error term at time t.
- ϵ_{t-j}^2: The squared error terms from previous periods $(t - j)$.
- σ_{t-j}^2: The conditional variance from previous periods $(t - j)$.
- a_0: A constant term ensuring the positivity of σ_t^2.
- a_j: Coefficients for the lagged squared error terms, representing the ARCH component.
- b_j: Coefficients for the lagged conditional variances, representing the GARCH component.
- p: The number of lagged variances included in the model.
- q: The number of lagged squared error terms included in the model.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The inclusion of the GARCH component ($b_j \sigma_{t-j}^2$) allows the model to efficiently capture volatility clustering, a characteristic feature of financial time series, where periods of high volatility tend to be followed by high volatility, and periods of low volatility tend to cluster together.

By combining the ARCH and GARCH components, the GARCH model provides a more parsimonious representation of volatility dynamics, making it a popular choice in financial applications such as risk management, portfolio optimization, and option pricing. The GARCH(1,1) model, where $p = 1$ and $q = 1$, is particularly widely used due to its simplicity and effectiveness in capturing most volatility patterns in practice.

3.6 Exponential GARCH Model (EGARCH)

The Exponential GARCH (EGARCH) model, introduced by Nelson (1991), is a variation of the GARCH family designed to address some limitations of traditional GARCH models. Unlike standard GARCH models, which assume that volatility is determined solely by the magnitude of shocks and not their direction, the EGARCH model allows for asymmetry, meaning that positive and negative shocks can have different impacts on volatility. This feature is particularly relevant in financial markets, where bad news often causes more significant volatility than good news of similar magnitude.

Another distinguishing characteristic of the EGARCH model is that it directly models the logarithm of the conditional variance. This approach ensures that the conditional variance is always positive without requiring non-negative coefficients, as is necessary in traditional GARCH models.

The EGARCH(p, q) model is expressed as:

$$\log(\sigma_t^2) = a_0 + \sum_{j=1}^{q} a_j g(Z_{t-j}) + \sum_{j=1}^{p} b_j \log(\sigma_{t-j}^2) \tag{3.8}$$

- σ_t^2: The conditional variance at time t.
- Z_{t-j}: The standardized error term at time $t - j$, calculated as $Z_t = \frac{\epsilon_t}{\sigma_t}$.
- $g(Z_{t-j})$: A function that incorporates asymmetry, given by:

$$g(Z_{t-j}) = \theta Z_{t-j} + \gamma(|Z_{t-j}| - E[|Z_{t-j}|])$$

- θ: Parameter measuring the impact of the sign of shocks (asymmetry effect).
- γ: Parameter capturing the magnitude of shocks relative to their expected value.
- a_0: Constant term.
- a_j: Coefficients for the asymmetric effects of shocks.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- b_j: Coefficients for the lagged logarithms of the conditional variance.
- p: Number of lagged conditional variances included.
- q: Number of lagged error terms included.

By modeling the logarithm of the variance, the EGARCH model allows for greater flexibility and ensures that the conditional variance remains positive without imposing restrictive non-negativity constraints on the coefficients. The asymmetry captured through $g(Z_{t-j})$ accounts for the fact that negative shocks (e.g., bad news) may lead to higher volatility than positive shocks (e.g., good news) of the same size.

This makes the EGARCH model particularly useful in financial and economic applications where asymmetric responses to news are common. Its ability to handle volatility clustering, asymmetry, and long memory effects in a more parsimonious framework has made it a widely adopted tool for analyzing time-varying volatility.

3.7 Integrated GARCH Model (IGARCH)

The Integrated Generalized Autoregressive Conditional Heteroskedasticity (IGARCH) model represents a significant refinement in the analysis of financial time series, particularly in modeling the persistence of volatility shocks, Engle and Bollerslev (1986). While the standard GARCH model assumes that the influence of past shocks gradually diminishes over time, the IGARCH model introduces a key distinction: the impact of volatility shocks is assumed to be permanent. This is achieved by constraining the sum of the autoregressive and moving average coefficients to equal one, ensuring that the effects of past shocks on volatility do not fade over time.

The IGARCH(p, q) model is expressed as:

$$\sigma_t^2 = a_0 + \sum_{j=1}^{q} a_j \epsilon_{t-j}^2 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2 \tag{3.9}$$

subject to the constraint:

$$\sum_{j=1}^{q} a_j + \sum_{j=1}^{p} b_j = 1 \tag{3.10}$$

- σ_t^2: The conditional variance of the error term at time t.
- ϵ_{t-j}^2: The squared error terms from previous periods $(t - j)$.
- a_0: A constant term ensuring positivity of σ_t^2.
- a_j: Coefficients representing the influence of past squared error terms.
- b_j: Coefficients representing the influence of past conditional variances.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- p: The number of lagged conditional variances included in the model.
- q: The number of lagged squared error terms included in the model.

A distinctive characteristic of the IGARCH model is its implication for the unconditional variance. Under the unit root condition ($\sum a_j + \sum b_j = 1$), the unconditional variance becomes undefined. This property presents challenges for theoretical analysis and practical applications, particularly in areas like risk management and derivative pricing, where accurate volatility forecasts are essential.

Advantages and Limitations

The IGARCH model is especially useful for financial time series exhibiting long memory in volatility, where the effects of past shocks remain significant over extended periods. Its ability to model high persistence in volatility has been widely acknowledged, with empirical studies often demonstrating its superior fit for series characterized by prolonged volatility clustering.

However, the assumption of permanent volatility shocks can be restrictive in some contexts. This limitation has spurred the development of alternative models, such as fractional GARCH, which provide greater flexibility in capturing long memory while allowing for a gradual decay in the impact of past shocks.

Despite these challenges, the IGARCH model remains a valuable tool in financial econometrics for understanding and forecasting time series with persistent volatility.

3.8 GJR-GARCH and TGARCH Models

The Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) model and the Threshold GARCH (TGARCH) model are two prominent extensions of the traditional GARCH framework that address the asymmetric nature of volatility observed in financial time series. Both models incorporate additional terms into the conditional variance equation to account for the differing impacts of positive and negative shocks, a phenomenon often referred to as the leverage effect.

The conditional variance equation for both the GJR-GARCH and TGARCH models can be expressed as:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2 + \sum_{k=1}^{q} \gamma_k \epsilon_{t-k}^2 I_{t-k} \tag{3.11}$$

- σ_t^2: The conditional variance at time t.
- ϵ_{t-i}^2: The squared error terms from previous periods $(t-i)$.
- σ_{t-j}^2: The lagged conditional variances $(t-j)$.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- I_{t-k}: An indicator function that equals 1 if $\epsilon_{t-k} < 0$ (negative shocks) and 0 otherwise.
- a_0: A constant term ensuring positivity of σ_t^2.
- a_i: Coefficients for the lagged squared error terms, representing the ARCH component.
- b_j: Coefficients for the lagged conditional variances, representing the GARCH component.
- γ_k: Coefficients capturing the asymmetric effect of negative shocks on volatility.
- p: The number of lagged conditional variances included in the model.
- q: The number of lagged squared error terms included in the model.

Although the mathematical structure of the two models is similar, their interpretation and focus differ slightly:

- **GJR-GARCH Model**: Developed by Glosten et al. (1993), this model effectively captures the asymmetry in volatility responses. It highlights how negative shocks, such as unexpected declines in asset prices, can lead to greater increases in volatility compared to positive shocks of the same magnitude. The inclusion of the $\gamma_k \epsilon_{t-k}^2 I_{t-k}$ term reflects this phenomenon, known as the leverage effect, making the GJR-GARCH model widely used in financial market analysis.
- **TGARCH Model**: Proposed by Zakoian (1994) and also addressed by Glosten et al. (1993), the TGARCH model employs a threshold mechanism to differentiate the impacts of positive and negative shocks. Using the indicator function I_{t-k}, it determines the regime based on the sign of the shock, allowing for a nuanced analysis of volatility behavior. This flexibility makes TGARCH particularly useful in modeling leverage effects and asymmetric volatility patterns under diverse market conditions.

3.9 Generalized Error Distribution (GED)

The Generalized Error Distribution (GED) is a flexible probability distribution often used in econometric models to account for non-normality in the residuals, Nelson(1991). Unlike the normal distribution, the GED includes an additional shape parameter, ν, which controls the kurtosis of the distribution:

- For $\nu = 2$, the GED reduces to the standard normal distribution.
- For $\nu < 2$, the GED exhibits heavier tails than the normal distribution, capturing excess kurtosis often observed in financial time series.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- For $\nu > 2$, the distribution has thinner tails than the normal distribution.

This flexibility makes the GED particularly useful in modeling financial time series with extreme events or high kurtosis, where the standard assumption of normally distributed residuals is inadequate. It is frequently applied in GARCH models to better capture the properties of financial market returns.

3.10 Structural Breaks Analysis

Structural breaks analysis is a technique used to identify changes in the underlying data-generating process of a time series. These changes, often referred to as breakpoints, can result from shifts in economic policies, financial crises, technological advancements, or other external events that alter the behavior of the series.

The presence of structural breaks can significantly impact model accuracy and inference. Ignoring such breaks may lead to biased parameter estimates and incorrect conclusions. Common methods for detecting structural breaks include the Bai-Perron test and the Chow test, which identify multiple breakpoints in a time series.

In financial time series, structural breaks often correspond to major market events, such as recessions or regulatory changes. Accounting for these breaks allows for improved model performance and more accurate forecasting.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 4

Forecasting Approach

4.1 Forecasting with Rolling Window Approach

The rolling window approach is a common technique in time series forecasting, allowing models to adapt to changing patterns over time, as discussed by Tashman (2000). This method updates the training dataset by incorporating the most recent observations, enabling the model to better reflect shifts in the underlying data. It is especially useful in dynamic settings where the behavior of the data evolves frequently.

Overview of the Rolling Window Approach

The rolling window framework uses a rolling subset of the data for training in order to make forecasts iteratively. The model is updated with the most recent data at each stage, and a forecast for the desired time frame is created. In order to keep the model responsive to current trends, the window "rolls" forward by removing the oldest observation and adding the most recent one.

Steps in the Rolling Window Forecasting Process

The forecasting process using the rolling window approach involves the following steps:

1. **Dataset Division:** The data is split into a training set for building the model and a test set for evaluation. The training set starts with a fixed number of observations, while the test set contains the periods to be forecasted.

2. **Model Fitting:** The model is trained on the observations within the current window.

3. **Prediction:** Forecasts are generated for the next time point in the test set based on the trained model.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

4. **Window Adjustment:** The window shifts forward by adding the latest observation and dropping the oldest one. This process repeats until forecasts are made for all test periods.

Advantages of the Rolling Window Approach

The rolling window approach has lots of significant benefits. It makes the model extremely successful for studying non-stationary time series by enabling it to adjust to recent changes in the data. It improves predicted accuracy by focusing on the most recent observations, which lowers the possibility that out-of-date data could bias the findings. Because the window size may be changed to balance highlighting short-term fluctuations and recording long-term patterns, the method is also very adaptable. Furthermore, its iterative architecture ensures realistic and useful results by reflecting real-world forecasting processes, where predictions are improved as new data becomes available.

Applications of the Rolling Window Approach

The rolling window technique is widely used in many different fields. It is used in financial markets to predict volatility, returns, or stock prices while adjusting to the constantly shifting market conditions. It aids in forecasting important macroeconomic variables including GDP growth, inflation, and unemployment rates. The technique is also used in energy markets to model and predict important factors like supply and demand as well as the prices of products like electricity and oil. Because of its adaptability, it is a useful tool for working with dynamic and changing datasets in a variety of sectors. This approach offers a strong foundation for time series forecasting, guaranteeing that models continue to be applicable and efficient in identifying changing trends in data.

4.2 Forecasting with Random Forest Model

Because of its adaptability, capability to manage non-linear relationships, and resistance to overfitting, the Random Forest model is a flexible machine learning technique that has been used widely in predicting applications. The Random Forest model, which is based on the ensemble learning principle, combines the outputs of several decision trees to generate predictions, increasing stability and accuracy.

Overview of the Random Forest Model

A random subset of the dataset and a random subset of predictor variables are used to train each of the many decision trees that make up Random Forest. The model forecasts the target variable by selecting (in classification tasks) or averaging the outputs (in regression tasks) from each individual tree. This ensemble method improves the model's performance and lowers prediction variance.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Steps in Forecasting with Random Forest

The process of using the Random Forest model for forecasting involves the following steps:

1. **Data Preparation:** Organize the dataset into predictor variables (X) and the target variable (y), ensuring proper handling of time series features like lagged values or seasonality components.

2. **Feature Engineering:** Create additional predictors such as lagged variables, rolling averages, or categorical indicators to capture temporal patterns in the data.

3. **Model Training:** Train the Random Forest model using a training dataset. The model constructs multiple decision trees by splitting the dataset randomly and selecting random subsets of predictors at each node.

4. **Forecasting:** Use the trained model to predict the target variable for the test dataset or future periods.

5. **Evaluation:** Assess the model's performance using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE).

Advantages of the Random Forest Model for Forecasting

For predicting problems, the Random Forest model offers a number of benefits. It works well with complicated datasets because it captures the complex, non-linear interactions between predictors and the target variable. Predictions are more reliable because of the model's ensemble structure, which guarantees resistance to noise and outliers. Additionally, by emphasizing the relative contributions of predictors, Random Forests provide insights into feature relevance, facilitating improved interpretation. Another important aspect of the model is its adaptability to a variety of forecasting issues, as it can handle both numerical and categorical inputs.

Applications of Random Forest in Forecasting

Because of its adaptability and predictive ability, Random Forest has been used extensively in many different fields. It is employed in financial forecasting to help analysts navigate complex market dynamics by predicting stock prices, returns, and volatility. By projecting future demands for goods or services in industries like manufacturing and retail, it also plays a significant part in demand forecasting. The issues of shifting supply and demand are addressed in the energy markets by using Random Forest models to forecast electricity generation, consumption, or prices. It is also useful in macroeconomic analysis, where it aids in modeling patterns in important metrics like inflation or GDP growth.Nevertheless, Random Forest's incapacity to naturally account for data

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

reliance is one of its drawbacks when it comes to time series forecasting. To address this, methods such as block Random Forest combined with block bootstrap can be applied to better capture the time series structure and improve forecasting accuracy.

4.3 Forecasting with Support Vector Regression (SVR) Model

A machine learning method called Support Vector Regression (SVR) is founded on the ideas of Support Vector Machines (SVM). Because it can manage nonlinear interactions and produce reliable forecasts even when there is noise, it is a good fit for forecasting applications. SVR models are capable of efficiently capturing complex patterns in the data by employing kernel functions to translate input data to a higher-dimensional space.

Overview of the SVR Model

Finding a function that, within a given tolerance (ϵ), roughly represents the connection between the input predictors and the target variable is the aim of SVR. A loss function that penalizes departures from the actual target values beyond this margin is minimized by the model. SVR is very effective with sparse datasets because, in contrast to classical regression, it concentrates on the points close to the decision border (support vectors).

The general SVR objective function is:

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}[\xi_i + \xi_i^*] \tag{4.1}$$

subject to:

$$\begin{aligned} &y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i, \\ &(w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \geq 0. \end{aligned}$$

- w: Weight vector.
- b: Bias term.
- ξ_i, ξ_i^*: Slack variables for points outside the ϵ-tube.
- C: Regularization parameter controlling the trade-off between model complexity and error tolerance.
- ϵ: Insensitivity margin.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Steps in Forecasting with SVR

The process of using SVR for forecasting involves the following steps:

1. **Data Preparation:** Organize the dataset into predictor variables (X) and the target variable (y), and ensure appropriate preprocessing, such as scaling, to standardize input features.

2. **Feature Engineering:** Construct lagged variables, rolling statistics, or seasonality indicators to incorporate temporal patterns into the predictors.

3. **Model Selection:** Choose the kernel function (e.g., linear, polynomial, or radial basis function (RBF)) and tune hyperparameters (C, ϵ, kernel parameters) using cross-validation.

4. **Model Training:** Fit the SVR model to the training dataset, optimizing the loss function to minimize prediction errors.

5. **Forecasting:** Generate predictions for the target variable in the test dataset or future periods.

6. **Evaluation:** Assess model performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE).

Advantages of SVR for Forecasting

For forecasting applications, Support Vector Regression (SVR) has many benefits. By simulating complex dependencies between predictors and the target variable, it manages non-linear relationships with effectiveness. Because it can generalize effectively even in noisy datasets due to its insensitivity to slight deviations, the model is adaptable to noise. Another important strength is its flexibility. Its emphasis on support vectors makes it computationally economical and appropriate for large-scale issues.

Applications of SVR in Forecasting

Support Vector Regression (SVR) is frequently used for a variety of forecasting tasks because of its robustness against overfitting and capacity to catch non-linear patterns. It is employed in financial forecasting to forecast returns, volatility, and stock prices. By predicting future requirements in industries like manufacturing and retail, it also plays a significant part in demand prediction. SVR models are used in the energy markets to predict patterns in the use of gas, oil, and electricity. SVR is a useful tool for a variety of forecasting applications because of these advantages.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

4.4 Forecasting with Gradient Boosting Model

A powerful machine learning method for classification and regression applications, gradient boosting is renowned for its capacity to represent intricate, nonlinear relationships. Through a series of optimizations of a loss function, it constructs an ensemble of weak learners, usually decision trees. A highly accurate predictive model is produced by training each new tree to minimize the mistakes of the prior ensemble.

Overview of Gradient Boosting

Gradient Boosting's main concept is to iteratively improve the model by incorporating trees that fix the remaining mistakes from earlier iterations. Gradient Boosting trains trees progressively, concentrating on the most difficult observations, in contrast to Random Forest, which trains trees randomly.

The general form of the Gradient Boosting model is:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x), \tag{4.2}$$

- $F_m(x)$: The prediction function after m-th iteration.
- $F_{m-1}(x)$: The prediction function from the previous iteration.
- η: The learning rate, controlling the contribution of each tree.
- $h_m(x)$: The new tree trained to minimize the residual errors of $F_{m-1}(x)$.

Steps in Forecasting with Gradient Boosting

The forecasting process using Gradient Boosting involves the following steps:

1. **Data Preparation:** Split the dataset into training and test sets. Ensure the features are appropriately preprocessed, including handling missing values, encoding categorical variables, and scaling numerical features if needed.
2. **Feature Engineering:** Generate additional predictors, such as lagged variables, rolling means, or seasonality indicators, to capture temporal patterns in the data.
3. **Model Training:** Train the Gradient Boosting model on the training set, tuning hyperparameters such as the learning rate, number of trees, and tree depth using cross-validation.
4. **Forecasting:** Use the trained model to predict the target variable for the test set or future time periods.
5. **Evaluation:** Assess model performance with metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE).

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Advantages of Gradient Boosting for Forecasting

For forecasting applications, gradient boosting has many benefits. By iteratively improving the model and guaranteeing improved performance at every stage, it achieves great predicted accuracy. The approach is very adaptable, capturing nonlinear interactions between predictors and the target variable and allowing a wide range of loss functions. It also helps interpret the relative influence of predictors by offering insights into feature relevance. Another important advantage is its adaptability, which allows for the improving of hyperparameters like learning rate and tree depth to maximize performance for certain forecasting demands.

Applications of Gradient Boosting in Forecasting

Because of its ability to handle intricate patterns and its repeated optimization process, gradient boosting is widely employed in a variety of forecasting applications. It is used in financial markets to forecast stock prices, returns, and risk metrics, offering useful information for making decisions. The technique is crucial for distribution and retail management since it is also frequently used in demand forecasting to project future requirements for goods and services. Gradient Boosting models are used in the energy sector to predict the production of renewable energy sources and electricity consumption.

4.5 Forecasting Performance Metrics

To evaluate the predictive performance of the models, three commonly used error metrics are employed: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), Chai and Draxler (2014),Hyndman and Koehler (2006). These metrics provide complementary perspectives on forecasting accuracy, allowing for a robust assessment of model performance. A lower value for these metrics indicates better alignment between the predicted and actual values.

4.5.1 Mean Absolute Error (MAE)

The average size of prediction mistakes, independent of their direction, is measured by the Mean Absolute Error (MAE). It is described as:

$$MAE = \frac{1}{T} \sum_{t=1}^{T} |y_t - \hat{y}_t| \tag{4.3}$$

T represents the total number of observations, y_t is the actual value at time t, and $\hat{y}_t$ is the predicted value. MAE provides a straightforward interpretation, expressing the average error in the same units as the data.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

4.5.2 Mean Squared Error (MSE)

The Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values. It is computed as:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \tag{4.4}$$

This metric penalizes larger errors more heavily than smaller ones, making it particularly sensitive to outliers. MSE is widely used in applications where minimizing large deviations in predictions is critical.

4.5.3 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is derived from the MSE and provides a measure of error in the same units as the target variable. It is defined as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{4.5}$$

RMSE combines the interpretability of MAE with the sensitivity to larger errors inherent in MSE, making it a balanced metric for evaluating overall model performance.

Interpreting the Metrics

Each metric serves a distinct purpose:

- **MAE:** Measures the average error magnitude, providing a simple and intuitive summary of model accuracy.
- **MSE:** Highlights larger errors due to its quadratic formulation, which is useful when large deviations are particularly undesirable.
- **RMSE:** Combines the advantages of MAE and MSE by penalizing large errors while maintaining interpretability in the same units as the target variable.

It is uncommon for a single model to obtain the lowest error across all measurements because each metric has a unique focus. In order to ensure an accurate evaluation, the forecasting models' performance is assessed separately for every metric in this study. Based on the combined insights from all indicators, conclusions are made.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 5

Empirical Application

In our analysis, we draw on monthly pricing data for three of the most critical crude oil benchmarks: West Texas Intermediate (WTI), Brent crude, and Dubai crude. These benchmarks are key indicators of oil prices globally and provide a comprehensive view of the market.

WTI is well-regarded for its light and sweet qualities, making it a preferred choice in U.S. refineries. The data we use covers WTI prices from July 2001, where it started at 26.43 per barrel, up to October 2023, when it reached 85.64 per barrel. This dataset allows us to track the evolution of U.S. oil pricing over two decades, including significant economic shifts and market events.

Brent crude, sourced from the North Sea, is another major player in global oil markets. It's widely used to price oil in Europe, Africa, and the Middle East. Our dataset traces Brent prices from July 2001, beginning at 24.99 per barrel, through to October 2023, where the price stands at 88.95 per barrel. This long-term data provides insights into how Brent has influenced global oil markets over the years.

Dubai crude plays a crucial role in pricing for the Middle Eastern and Asian markets. The data we analyze tracks Dubai crude from July 2001, starting at 23.40 per barrel, and follows its trajectory to October 2023, when it was priced at 88.79 per barrel. This dataset is particularly valuable for understanding the oil market dynamics in these regions.

By incorporating data from these three benchmarks, our analysis captures a wide spectrum of market trends and economic factors affecting oil prices on a global scale. This approach ensures a well-rounded understanding of both regional and international energy markets.

5.1 Oil Price Returns Analysis

Building on the pricing data gathered for West Texas Intermediate (WTI), Brent crude, and Dubai crude, we proceed to analyze their respective returns. Returns provide insight into the relative price movements between consecutive time pe-

UNIVERSITY OF ECONOMICS & BUSINESS ΠΑΝΕΠΙΣΤΗΜΙΟ ΟΙΚΟΝΟΜΙΚΟ ΑΘΗΝΩΝ ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

riods, helping us understand market volatility and response to global events.

5.1.1 Calculation of Returns

The monthly returns for each crude oil benchmark are calculated using the formula for log returns:

$$\text{Log Returns} = \log\left(\frac{P_t}{P_{t-1}}\right)$$

P_t is the price at time t and P_{t-1} is the price at the previous time period. Log returns are commonly used in financial analysis as they provide a symmetrical measure of percentage changes and facilitate aggregation over multiple periods. This method allows for a clearer analysis of oil price volatility and the response to global events.

5.1.2 WTI Crude Oil Returns

The WTI crude oil returns (Figure 5.1) show distinct periods of heightened volatility. The sharp fluctuations during the 2008 financial crisis and the 2020 COVID-19 pandemic are particularly notable, indicating significant market disruptions.

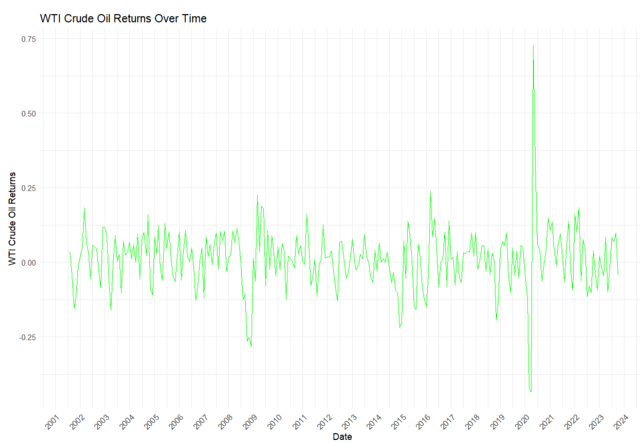

Figure 5.1: WTI Crude Oil Returns Over Time

5.1.3 Brent Crude Oil Returns

The Brent crude oil returns, as shown in Figure 5.2, reveal a similar pattern of volatility, particularly during the 2008 and 2020 market shocks. Brent, being a widely used global benchmark, closely mirrors the overall trends observed in international oil markets.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

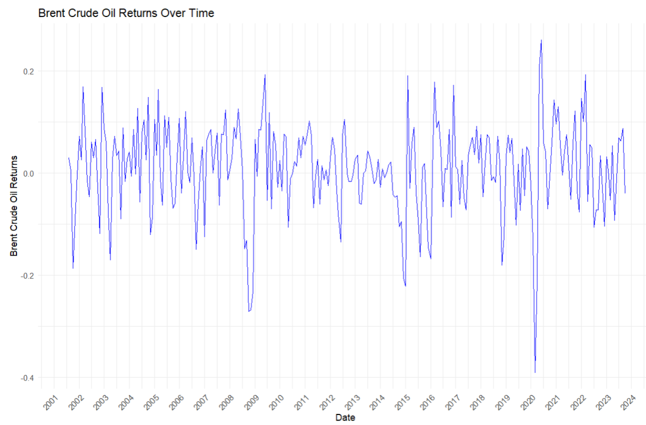

Figure 5.2: Brent Crude Oil Returns Over Time

5.1.4 Dubai Crude Oil Returns

Figure 5.3 shows the returns for Dubai crude oil, which reflect similar disruptions around the 2008 and 2020 periods. Given its relevance to Middle Eastern and Asian markets, the Dubai crude oil return trends offer insight into how these regions have been impacted by global events.

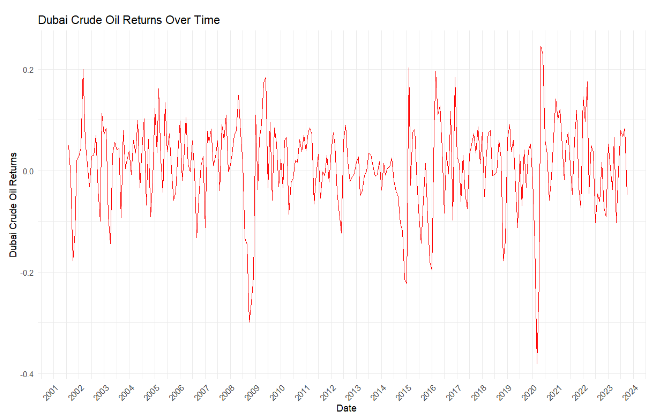

Figure 5.3: Dubai Crude Oil Returns Over Time

5.1.5 Summary

In summary, the returns analysis for WTI, Brent, and Dubai crude oils reveals a consistent pattern of volatility during major economic crises, such as the 2008 financial collapse and the COVID-19 pandemic. These plots highlight the sensitivity of global oil prices to macroeconomic events and provide a clearer

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

understanding of the interconnectedness of these benchmarks.

5.2 Objective of the Study

The primary objective of this analysis is to understand how various factors influence global crude oil prices, specifically for Brent, WTI, and Dubai crude benchmarks. The dataset used for this purpose includes a comprehensive set of variables that could potentially affect these oil prices.

The analysis contains both the dependent variables (the crude oil prices) and a range of explanatory variables. Additionally, there are separate files for each oil price that follow a specific structure: the first column corresponds to the crude oil price (y), while the remaining columns represent the potential influencing factors (x).

5.2.1 Dependent Variables

The following global crude oil returns are considered as the dependent variables:

- **Global price of WTI Crude**: The price per barrel of West Texas Intermediate crude, a major benchmark in the U.S. market.
- **Global price of Brent Crude**: The price per barrel of Brent crude, which is widely used as a benchmark in Europe, Africa, and the Middle East.
- **Global price of Dubai Crude**: The price per barrel of Dubai crude, often used in Middle Eastern and Asian markets.

5.2.2 Explanatory Variables

Several macroeconomic indicators, market sentiment indices, and interest rate data are considered as potential drivers of oil prices. These include:

- **Equity Market Volatility Tracker: Overall**: A measure of overall market volatility, capturing investor sentiment and market stability.
- **Economic Policy Uncertainty Index for Europe**: This index tracks the uncertainty in economic policies within Europe, which may affect investment and oil demand.
- **Economic Policy Uncertainty Index for United States**: Similar to the European index, but focused on the U.S., reflecting uncertainty in fiscal and monetary policies.
- **Current General Business Conditions; Diffusion Index for New York**: A diffusion index that measures the general business climate in New York, often used as an economic indicator.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- **Federal Funds Effective Rate**: The interest rate at which depository institutions lend balances to each other overnight, influencing broader financial conditions.
- **3-Month Treasury Bill Secondary Market Rate, Discount Basis**: The rate for 3-month Treasury bills, representing short-term risk-free interest rates.
- **6-Month Treasury Bill Secondary Market Rate, Discount Basis (TB6MS)**: Similar to the 3-month Treasury bill rate but for a 6-month period.
- **Equity Market Volatility: Infectious Disease Tracker**: A tracker measuring volatility in equity markets specifically tied to infectious disease outbreaks, a factor that can impact global oil demand.
- **Global price of Nickel**:The price of nickel, another globally traded commodity, whose price fluctuations might correlate with crude oil markets.

The aim is to assess the extent to which these factors influence the price movements of WTI, Brent, and Dubai crude oil, and to determine which of these explanatory variables have the most significant impact on each oil price.

5.2.3 Stationarity Check: Augmented Dickey-Fuller Test

To ensure that the returns series for WTI crude oil is suitable for further analysis, we conducted the Augmented Dickey-Fuller (ADF) test, Dickey and Fuller (1979), to check for stationarity. The hypotheses for the test are as follows:

- $\mathbf{H_0}$: The series is non-stationary.
- $\mathbf{H_1}$: The series is stationary.

The test was performed at three significance levels: $\alpha = 1\%$, $\alpha = 5\%$, and $\alpha = 10\%$. The decision rule is as follows:

- If $\alpha >$ p-value, we reject the null hypothesis H_0 and conclude that the series is stationary at the chosen significance level.
- If $\alpha <$ p-value, we fail to reject the null hypothesis H_0 and conclude that the series is non-stationary.

For the WTI returns series, the result of the ADF test is as follows:

p-value: 0.01

Since the p-value (0.01) is less than the significance level $\alpha = 1\%$, we reject the null hypothesis H_0 and conclude that the WTI returns series is stationary at the 1% significance level. This indicates that the series does not exhibit a

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

unit root and is suitable for further analysis. Similarly, the p-value for Brent returns is less than $\alpha = 1\%$, leading us to reject H_0 and conclude that the Brent returns series is stationary. For the Dubai returns series, the p-value is also less than $\alpha = 1\%$, indicating that the series is stationary.

Based on the results of the ADF test, we conclude that the returns series for WTI, Brent, and Dubai crude oil are all stationary at the 1% significance level. This means that the series do not exhibit a unit root and are suitable for further time series analysis.

5.3 Analysis of WTI Crude Oil Returns

5.3.1 Multiple Regression for WTI Returns

Multiple regression models can incorporate several explanatory variables. The general form of the model is:

$$Y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \epsilon_t$$
$$\epsilon_t \sim N(0, \sigma^2)$$

where $X_{1,t}, X_{2,t}, \ldots, X_{k,t}$ represent the explanatory factors, and ϵ_t is the error term, assumed to follow a normal distribution with a mean of 0 and variance σ^2. To assess the impact of various macroeconomic and market variables on the returns of WTI crude oil, we performed a multiple regression analysis. All the available explanatory variables were initially considered for the model, with WTI crude oil returns as the dependent variable.

Results of the Full Model

The regression analysis with all the explanatory variables produced the following results:

Table 5.1: Estimated Coefficients for Full Model of WTI Returns

Variable	**Estimate**	**p-value**
Intercept	0.02683	0.4323
EMVOVERALLEMV	-0.00377	0.0006***
EUEPUINDXM	-0.00021	0.0556.
USEPUINDXM	0.00062	0.0178*
GACDISA066MSFRBNY	0.00091	0.0494*
FEDFUNDS	0.00836	0.7995
TB3MS	0.00777	0.9140
TB6MS	-0.01282	0.8184
INFECTDISEMVTRACK	-0.00048	0.6837
PNICKUSDM	0.0000003149	0.7480
Adjusted R-Squared	**0.0799**	

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The full model explains approximately 7.99% of the variance in WTI crude oil returns, as indicated by the adjusted R-squared value. The results highlight that the following variables are statistically significant predictors at conventional significance levels: Equity Market Volatility Tracker, Economic Policy Uncertainty Index for Europe, Economic Policy Uncertainty Index for the United States and Current General Business Conditions. The remaining variables do not exhibit statistically significant effects.

Interpretation of Significant Variables

The analysis shows that overall market volatility, as measured by the Equity Market Volatility Tracker, has a negative and significant impact on WTI returns, suggesting that higher volatility reduces returns. Similarly, the Economic Policy Uncertainty Index for Europe negatively affects WTI returns, though its significance is marginal. In contrast, the Economic Policy Uncertainty Index for the United States has a positive influence, indicating that increased policy uncertainty in the U.S. is associated with higher returns. Additionally, the Current General Business Conditions Index for New York positively impacts WTI returns, reflecting the effect of improved business conditions. Other variables, such as interest rates, infectious disease market volatility, and nickel prices, do not significantly influence WTI returns, implying their effects may be minimal or captured indirectly through other factors.

Diagnostic Tests for the Residuals

Several diagnostic tests were performed to check the assumptions of the regression model:

Autocorrelation of the Residuals

The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the residuals in Figure 5.4 showed significant autocorrelation at lag 1.

Autocorrelation of the squared Residuals

The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the squared residuals in Figure 5.5 showed significant autocorrelation at lags 1, 2, and 3 suggesting the presence of volatility clustering in the residuals.

Normality Test

The Jarque-Bera test and Shapiro-Wilk test were used to test the normality of the residuals. Both tests rejected the null hypothesis of normality with p-values less than 0.05, indicating that the residuals do not follow a normal distribution. So we will continue using t-distribution for the residuals.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

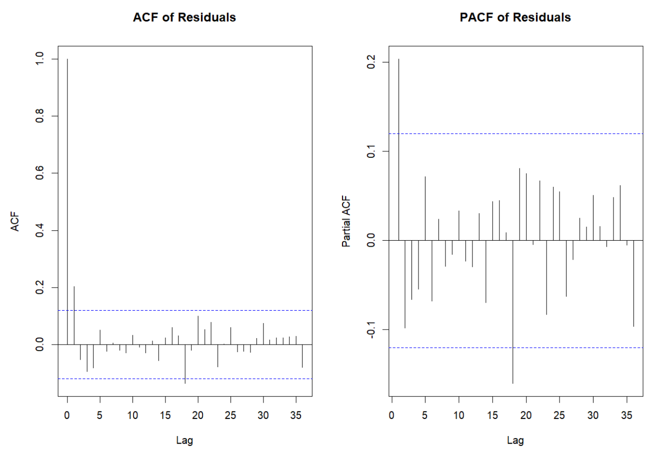

Figure 5.4: ACF and PACF of the WTI Residuals

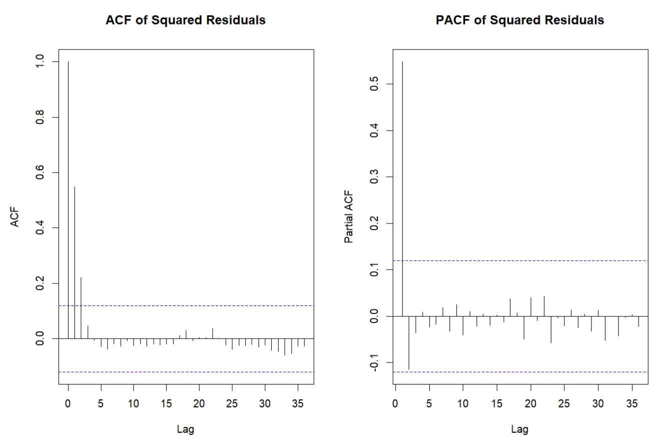

Figure 5.5: ACF and PACF of the WTI Squared Residuals

5.3.2 ARMA-GARCH Model Estimation

To better capture the volatility clustering observed in the residuals of the regression model, we fitted a series of ARMA(1,1)-GARCH models, starting with a standard GARCH(1,1) specification.

The model was constructed using all the explanatory variables as external regressors. It is worth noting that while different values of p and q were tested in the models, for simplicity and clarity, we present only the results of the ARMA(1,1)-GARCH(1,1) models. These provide a representative example of the findings.

Below are the steps and diagnostic tests performed:

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

ARMA(1,1)-GARCH(1,1) Model

The ARMA(1,1)-GARCH(1,1) model was initially fitted using a standard normal distribution for the residuals. The Ljung-Box test on the residuals returned a p-value of 0.6581, suggesting no significant autocorrelation at lag 36. However, the Ljung-Box test on the squared residuals gave a p-value less than 0.05, indicating the presence of autocorrelation in the squared residuals and potential ARCH effects that the model could not capture. The results are confirmed at the next Figures.

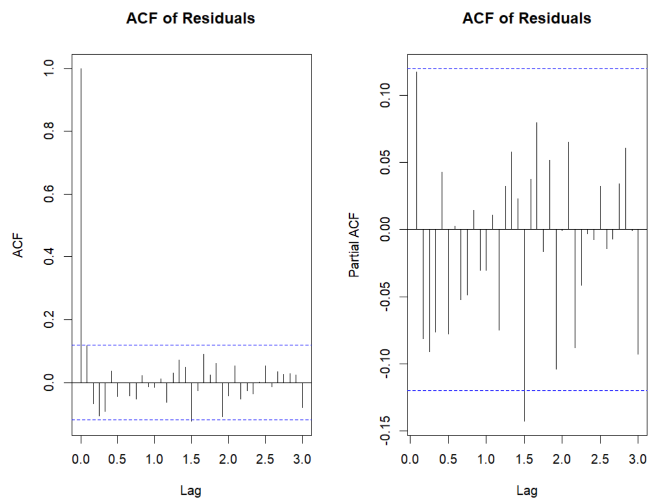

Figure 5.6: ACF and PACF of Residuals ARMA(1,1)- GARCH(1,1)

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

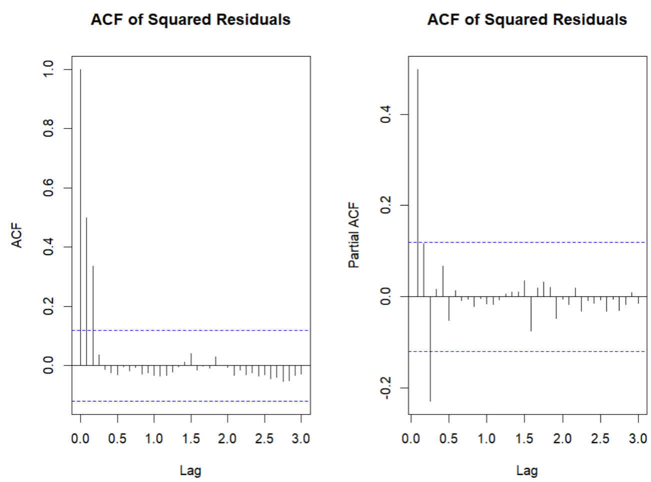

Figure 5.7: ACF and PACF of squared Residuals ARMA(1,1)- GARCH(1,1)

To address this, we proceeded to test more complex GARCH models.

ARMA(1,1)-EGARCH Model

In order to account for possible asymmetry in the volatility, an exponential GARCH (EGARCH) model was also taken into consideration. To take into consideration potential fat tails, the model employed a Student's t-distribution for the residuals. Once more, the squared residuals revealed persistent autocorrelation, indicating that more refining was required, even if the residuals showed no autocorrelation.

ARMA(1,1)- TGARCH and GARCH with GED

Additionally, a GARCH model that assumed a Generalized Error Distribution (GED) for the residuals was examined, as was a TGARCH model. The squared residuals for both models continued to show autocorrelation, indicating that volatility clustering was still a problem.

ARMA(1,1)-IGARCH Model

The next model to be examined was an Integrated GARCH (IGARCH). The residuals of this model's diagnostic tests showed no significant autocorrelation. The Box-Ljung test of the squared residuals, however, shows that volatility clustering remained and that the model was unable to fully account for the heteroscedasticity in the data.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

ARMA(1,1)-GJR-GARCH Model

Glosten-Jagannathan-Runkle GARCH (GJR-GARCH) was the last model examined. It introduces an asymmetric effect in which volatility is more affected by negative shocks than by positive ones. The mean dynamics were confirmed to be well defined by the diagnostic testing, which showed no significant autocorrelation in the residuals. Nevertheless, the Box-Ljung test indicates that volatility clustering continued even after asymmetry was incorporated into the model.

Structural Breaks Analysis

To account for potential structural breaks in the time series of WTI returns, we applied the Bai-Perron structural break test. This test is designed to detect multiple breakpoints in time series data. Using the breakpoints method, we identified significant breakpoints in the data.

The optimal number of breakpoints was determined based on the Residual Sum of Squares (RSS). There are four structural breaks in the WTI returns data.

The detected breakpoints are as follows:

- **Breakpoint 1:** 2005 (August)
- **Breakpoint 2:** 2009 (January)
- **Breakpoint 3:** 2014 (May)
- **Breakpoint 4:** 2020 (March)

These breakpoints correspond to significant periods in the oil market, such as the 2008 financial crisis, oil price shocks, and the COVID-19 pandemic in 2020. These structural breaks are important to consider when modeling the volatility of WTI returns, as they may indicate significant shifts or patterns in the data.

ARMA(1,1)-GARCH Model with Structural Breaks

To account for the detected structural breaks, we incorporated them as external regressors in an ARMA(1,1)-GARCH(1,1) model. The model used a Student's t-distribution for the residuals to handle potential fat tails. The external regressors were the detected breakpoints from the structural break analysis, allowing for shifts in volatility dynamics.

The Ljung-Box test was performed to check for autocorrelation in both the residuals and the squared residuals of the model. The results are shown in Table 5.2.

As shown in Table 5.2, both the residuals and the squared residuals passed the Ljung-Box test, with p-values above 0.05. This suggests that there is no significant autocorrelation remaining in either the residuals or their squares, indicating that the model adequately captures the volatility dynamics.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Table 5.2: Ljung-Box Test Results for Residuals and Squared Residuals

Test	**p-value**
Ljung-Box Test (Residuals)	0.57
Ljung-Box Test (Squared Residuals)	0.91

Model's Equation

The equation of the model is the following:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \sum_{i=1}^{9} \beta_i X_{t,i} + \sum_{j=1}^{4} \gamma_j \text{Break}_j + \epsilon_t$$

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

Structural Breaks

- Dummy variables:

$$\text{Break}_j = \begin{cases} 1 & \text{if } t > \text{breakpoint}_j, \\ 0 & \text{otherwise.} \end{cases}$$

The dummies adjust the intercept via γ_j to account for regime changes.

Ljung-Box Test Results for Various GARCH Models

For each of the GARCH models tested—ARMA(1,1)-EGARCH(1,1), ARMA(1,1)-GARCH(1,1) with GED, ARMA(1,1)-IGARCH(1,1), and ARMA(1,1)-GJR-GARCH(1,1)—we conducted the Ljung-Box test on both the residuals and squared residuals. The p-values for each model are reported in Table 5.3.

Table 5.3: Ljung-Box Test Results for Various GARCH Models (p-values)

Model	**Residuals (p-value)**	**Squared Residuals (p-value)**
ARMA(1,1)-EGARCH(1,1)	0.6977	0.7048
ARMA(1,1)-GARCH with GED	0.57	0.91
ARMA(1,1)-IGARCH(1,1)	0.57	0.97
ARMA(1,1)-GJR-GARCH(1,1)	0.6736	0.8315

All the p-values are above typical significance levels (such as 0.05), indicating that there is no significant autocorrelation left in either the residuals or the squared residuals for any of the models. This suggests that the models have adequately captured the volatility structure in the data.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

5.3.3 Models Comparison

In order to select the best model, the above models were fitted, and their performance was evaluated based on information criteria. The models considered were:

- GARCH(1,1)
- EGARCH(1,1)
- GJR-GARCH(1,1)
- IGARCH(1,1)
- GARCH(1,1) with GED distribution

The following information criteria were used to compare the models:

- **Akaike Information Criterion (AIC)**
- **Bayesian Information Criterion (BIC)**
- **Hannan-Quinn Information Criterion (HQIC)**

The table below shows the values of these criteria for each model:

Table 5.4: Information Criteria for Model Comparison

Model	**AIC**	**BIC**	**HQIC**
GARCH(1,1)	-1.5880	-1.3193	-1.4801
EGARCH(1,1)	-2.0612	-1.7791	-1.9479
GJR-GARCH(1,1)	-2.0585	-1.7764	-1.9452
IGARCH(1,1)	-1.8764	-1.6211	-1.7739
GARCH(1,1) with GED	-1.4553	-1.1866	-1.3474

Interpretation of Results

The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC) are commonly used to compare models. In this case:

- The model with the lowest AIC is considered to provide the best fit to the data.
- The BIC introduces a stronger penalty for additional parameters, making it more conservative in selecting models.
- HQIC is similar to BIC but provides a different penalty.

Based on the results in Table 5.4, the **EGARCH(1,1)** model has the lowest AIC (-2.0612), BIC (-1.7791), and HQIC (-1.9479). Therefore, it is selected as the best model to capture the characteristics of the data.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

5.3.4 Variable Selection Process

To determine the best model, the initial specification included 9 external variables in the mean equation. A stepwise procedure was employed, where the variables were removed one by one based on their statistical significance and contribution to the model fit. At each step, diagnostic tests for the residuals were conducted to ensure that the model adequately captured the characteristics of the data.

This iterative process allowed for the identification of the optimal subset of variables that best captures the characteristics of the data while maintaining model parsimony.

The final model includes the variables: Equity Market Volatility Tracker, Economic Policy Uncer- tainty Index for Europe, Economic Policy Uncertainty Index for the United States, Current General Business Conditions, Equity Market Volatility: Infectious Disease Tracker and Global price of Nickel.

We applied an ARMA(1,1) model and various GARCH models to the residuals of the data. The results of the Box-Ljung test showed no significant autocorrelation in the residuals, indicating that the mean model was well-specified. However, the Box-Ljung test applied to the squared residuals revealed that volatility clustering remained in the data. This indicates that the model was unable to fully account for the heteroscedasticity present in the series. So we add structural breaks like previously and the results of diagnostic cheks are presented in following table

To address the remaining volatility clustering, structural breaks were added to the model as previously described. The results of the diagnostic checks for the residuals and squared residuals, after incorporating structural breaks, are presented in the following table:

Table 5.5: Ljung-Box Test Results for Various GARCH Models (p-values)

Model	**Residuals (p-value)**	**Squared Residuals (p-value)**
ARMA(1,1)-GARCH(1,1)	0.7066	0.8738
ARMA(1,1)-EGARCH(1,1)	0.6115	0.1853
ARMA(1,1)-GJR-GARCH(1,1)	0.6070	0.6738
ARMA(1,1)-IGARCH(1,1)	0.7066	0.8738
ARMA(1,1)-GARCH with GED	0.7066	0.8738

We select the best model based on information criteria like previously. The table below shows the values of these criteria for each model.

Based on the results in Table 5.6, the GJR-GARCH(1,1) model has the lowest AIC (-2.0502), BIC (-1.8084), and HQIC (-1.9531). Therefore, it is selected as the best model to capture the characteristics of the data.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Table 5.6: Information Criteria for Model Comparison

Model	AIC	BIC	HQIC
GARCH(1,1)	-1.5817	-1.3533	-1.4900
EGARCH(1,1)	-2.0485	-1.8066	-1.9513
GJR-GARCH(1,1)	-2.0502	-1.8084	-1.9531
IGARCH(1,1)	-1.8753	-1.6603	-1.7889
GARCH(1,1) with GED	-1.4502	-1.2218	-1.3585

The equation of the final model is the following:

Mean Equation:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \sum_{i=1}^{6} \beta_i X_{t,i} + \sum_{j=1}^{4} \gamma_j \text{Break}_j + \epsilon_t$$

Variance Equation:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma_1 \epsilon_{t-1}^2 I(\epsilon_{t-1} < 0)$$

5.3.5 Random Forest Approach

To identify the most significant predictors in the dataset, a Random Forest model was applied. The dependent variable was modeled using all the explanatory variables. The variable importance was measured using the Mean Decrease in Accuracy, which quantifies how much the prediction accuracy decreases when a specific variable is excluded.

The top predictors in the model are: Equity Market Volatility, Economic Policy Uncertainty Index for Europe, Economic Policy Uncertainty Index for the United States, Current General Business Conditions, Federal Funds Effective Rate, as it is presented in Figure 5.8.

Decision Tree for Random Forest

To better understand the relationships captured by the Random Forest model, a decision tree was extracted and visualized. The Figure 5.9 presents the resulting decision tree, where each node represents a splitting condition based on one of the predictors, and the terminal nodes display the predicted value along with the proportion of data points in that leaf.

The first split occurs on Equity Market Volatility Tracker: Overall, indicating it is the most important variable. The split at 38.2 separates the dataset into two main groups based on high or low market volatility. For instances where Equity Market Volatility is lower than 38.2, the next important split is based on Equity Market Volatility: Infectious Disease Tracker, further refining the

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

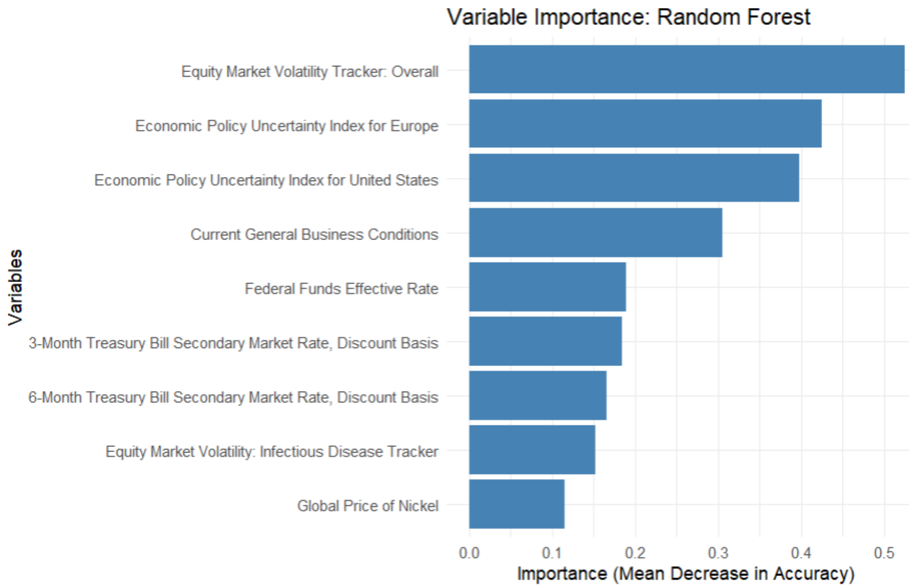

Figure 5.8: Variable Importance Plot from Random Forest

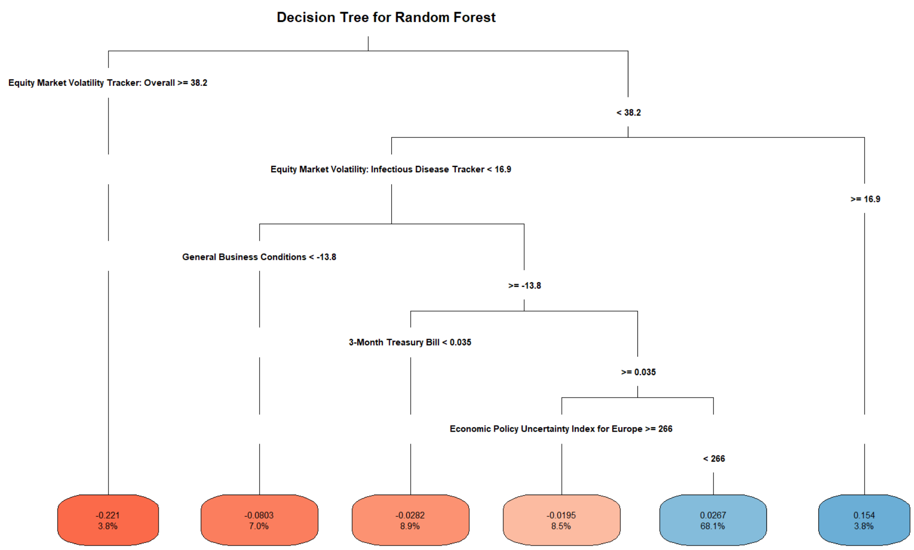

Figure 5.9: Decision Tree from Random Forest

decision-making process. For instances where Equity Market Volatility is bigger or equal than 38.2, the predicted value directly reflects a negative outcome (-0.221).

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The terminal nodes display the predicted values of WTI returns and the percentage of data points that fall into each leaf, providing insights into the distribution of outcomes.

5.4 Analysis of Brent Crude Oil Returns

5.4.1 Multiple Regression for Brent Returns

The same process as described for WTI returns was applied to Brent crude oil returns. The regression analysis with all explanatory variables produced the following results:

Table 5.7: Estimated Coefficients for Full Model of Brent Returns

Variable	**Estimate**	**p-value**
Intercept	0.03933	0.18347
EMVOVERALLEMV	-0.00322	0.00069***
EUEPUINDXM	-0.000198	0.03998*
USEPUINDXM	0.000395	0.08034.
GACDISA066MSFRBNY	0.000689	0.08506.
FEDFUNDS	-0.0103	0.71728
TB3MS	0.03858	0.53533
TB6MS	-0.02721	0.57298
INFECTDISEMVTRACK	-0.0002384	0.81371
PNICKUSDM	0.0000008831	0.29759
Adjusted R-Squared	**0.0979**	

The model explains approximately 9.79% of the variance in Brent crude oil returns, as indicated by the adjusted R-squared value. Significant predictors include: Equity Market Volatility Tracker, Economic Policy Uncertainty Index for Europe. Additionally, variables such as Economic Policy Uncertainty Index for the United States and Current General Business Conditions show significance at the 10% level.

5.4.2 Final Model for Brent Returns

Starting with the full model that included all explanatory variables (Table 5.7), we gradually removed variables that were not statistically significant. This step-by-step process helped us identify a simpler, more interpretable model that best captures the characteristics of the data. The aim was to retain only the variables with the strongest influence on Brent crude oil returns while ensuring the model remained robust and avoided overfitting. Diagnostic tests were conducted on the model as previously, to ensure its validity and robustness.

After several fit of various models we end up with the model without the variable Equity Market Volatility: Infectious Disease Tracker and 5 structural breaks based on RSS. The breakpoints are as follows:

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

- **Breakpoint 1:** 2005 (August)
- **Breakpoint 2:** 2008 (December)
- **Breakpoint 3:** 2012 (April)
- **Breakpoint 4:** 2016 (January)
- **Breakpoint 5:** 2020 (April)

The results of the diagnostic checks for the residuals and squared residuals, after incorporating structural breaks, are presented in the following table:

Table 5.8: Ljung-Box Test Results for Various GARCH Models (p-values) for Brent

Model	**Residuals (p-value)**	**Squared Residuals (p-value)**
ARMA(1,1)-GARCH(1,1)	0.1941	0.06828
ARMA(1,1)-EGARCH(1,1)	0.3171	0.08652
ARMA(1,1)-GJR-GARCH(1,1)	0.1899	0.04877
ARMA(1,1)-TGARCH(1,1)	0.2471	0.08269
ARMA(1,1)-IGARCH(1,1)	0.2108	0.06552
ARMA(1,1)-GARCH with GED	0.2758	0.3289

We select the best model based on information criteria like previously. The table below shows the values of these criteria for each model.

Table 5.9: Information Criteria for Model Comparison

Model	**AIC**	**BIC**	**HQIC**
GARCH(1,1)	-2.1781	-1.9094	-2.0702
EGARCH(1,1)	-2.2281	-1.9460	-2.1148
GJR-GARCH(1,1)	-2.1786	-1.8964	-2.0652
IGARCH(1,1)	-2.1775	-1.9222	-2.0749
GARCH(1,1) with GED	-1.7060	-1.4373	-1.5981

Based on the results in Table 5.9, the EGARCH(1,1) model has the lowest AIC (-2.2281), BIC (-1.9460), and HQIC (-2.1148). Therefore, it is selected as the best model to capture the characteristics of the data.

5.5 Analysis of Dubai Crude Oil Returns

5.5.1 Multiple Regression for Dubai Returns

The same process as described for WTI and Brent returns was applied to Dubai crude oil returns. The regression analysis with all explanatory variables produced the following results:

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Table 5.10: Estimated Coefficients for Full Model of Dubai Returns

Variable	**Estimate**	**p-value**
Intercept	0.04985	0.087739.
EMVOVERALLEMV	-0.00362	0.000118***
EUEPUINDXM	-0.0002018	0.033937*
USEPUINDXM	0.0003957	0.075871.
GACDISA066MSFRBNY	0.0005334	0.176258
FEDFUNDS	-0.006009	0.830416
TB3MS	0.03794	0.536611
TB6MS	-0.03011	0.527196
INFECTDISEMVTRACK	-0.0000708	0.943446
PNICKUSDM	0.0000007651	0.360157
Adjusted R-Squared	**0.1067**	

The model explains approximately 10.67% of the variance in Dubai crude oil returns, as indicated by the adjusted R-squared value. Significant predictors include: Equity Market Volatility Tracker, Economic Policy Uncertainty Index for Europe. Additionally, the variable Economic Policy Uncertainty Index for the United States demonstrates marginal significance at the 10% level.

5.5.2 Final Model for Dubai Returns

Starting with the full model that included all explanatory variables (Table 5.10), we gradually removed variables that were not statistically significant. This step-by-step process helped us identify a simpler, more interpretable model that best captures the characteristics of the data. The aim was to retain only the variables with the strongest influence on Dubai crude oil returns while ensuring the model remained robust and avoided overfitting. Diagnostic tests were conducted on the model as previously, to ensure its validity and robustness.

After several fit of various models we end up with the model Arma(1,1)-Igarch(1,1) without the variable Equity Market Volatility: Infectious Disease Tracker and Federal Funds Effective Rate.

The Ljung-Box test results for the residuals and squared residuals did not show significant autocorrelation, confirming the suitability of the model.

Table 5.11: Ljung-Box Test Results for Dubai Returns

Test	**p-value**
Ljung-Box Test (Residuals)	0.6331
Ljung-Box Test (Squared Residuals)	0.3229

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

5.6 Forecasting

5.6.1 Forecasting WTI Returns with Multiple Regression

To forecast the WTI returns for a 12-month period, we utilized a multiple regression model. We use the variables identified from the previous analysis as predictors: Equity Market Volatility Tracker, Economic Policy Uncer- tainty Index for Europe, Economic Policy Uncertainty Index for the United States, Current General Business Conditions, Equity Mar- ket Volatility: Infectious Disease Tracker and Global price of Nickel.

Model and Forecasting Process

The data was split into in-sample and out-of-sample datasets, where the last 12 observations were held as the out-of-sample data for validation. The multiple regression model was then fit to the in-sample data, and predictions were made for the out-of-sample period.

Performance Metrics

To evaluate the forecast accuracy, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as follows:

Table 5.12: Forecasting Performance Metrics for WTI Returns (Multiple Regression)

Metric	**Value**
MAE	0.0635
MSE	0.0055
RMSE	0.0746

These metrics indicate the average forecast error and provide a measure of the accuracy of our predictions.

Forecasting Results

The figure below shows the actual and predicted WTI returns for the last 12 months of the dataset:

In the plot, the blue line represents the actual WTI returns, while the red line represents the predicted values from the multiple regression model. As shown, the model captures some trends, though there are discrepancies in specific months, suggesting areas for potential model refinement.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

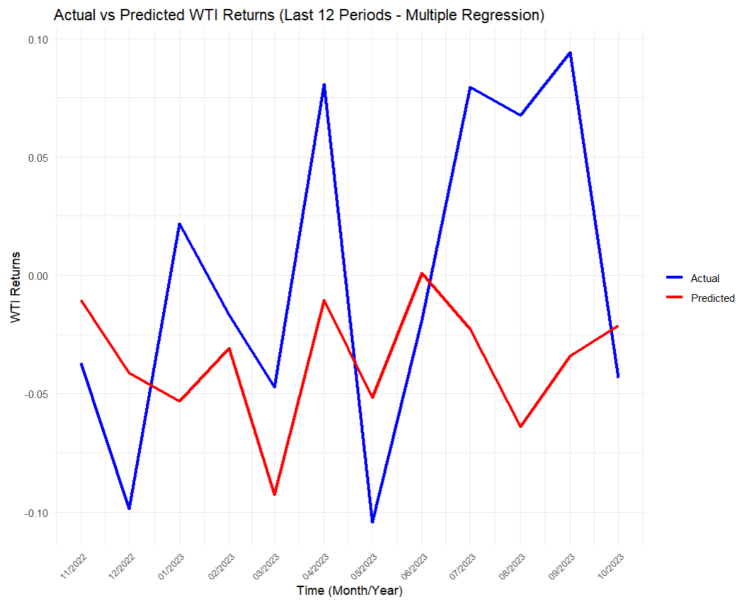

Figure 5.10: Actual vs Predicted WTI Returns (Last 12 Periods - Multiple Regression)

5.6.2 Forecasting WTI Returns with Rolling Window Approach

To improve forecast accuracy, we implemented a rolling window approach for predicting WTI returns over the next 12 months. This method continuously updates the training dataset with the most recent observations, allowing the model to adapt to new data patterns.

Performance Metrics

The following metrics were calculated to evaluate the forecast accuracy over the 12-month period:

Table 5.13: Forecasting Performance Metrics for WTI Returns (Rolling - Multiple Regression)

Metric	Value
MAE	0.0637
MSE	0.0054
RMSE	0.0737

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Forecasting Results

The figure below compares the actual WTI returns with the predicted values obtained using the rolling window approach.

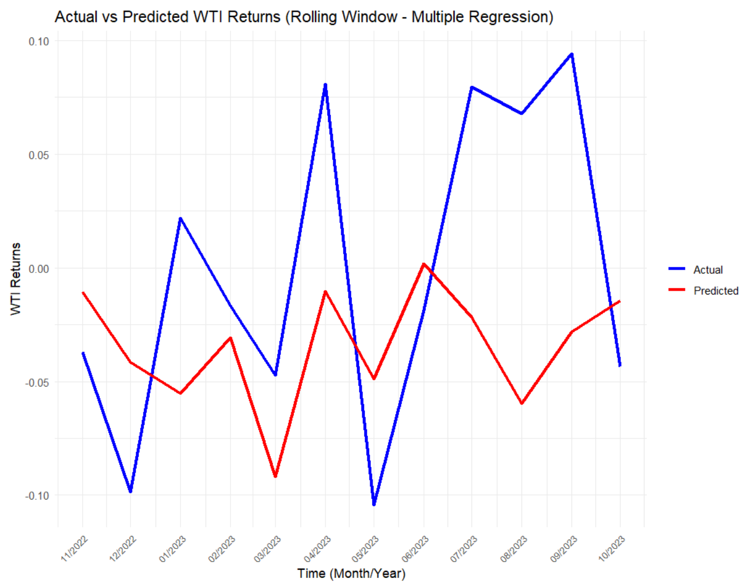

Figure 5.11: Actual vs Predicted WTI Returns (Rolling Window - Multiple Regression)

The plot illustrates the actual (blue) and predicted (red) WTI returns, showing how well the model adapts to recent trends over the 12-month forecasting horizon.

5.6.3 Forecasting WTI Returns with Regression Model Incorporating Structural Breaks

To forecast the WTI returns for a 12-month period, we applied a multiple regression model with additional dummy variables representing structural breaks as previously at the analysis.

Model and Forecasting Process

The data was split into in-sample and out-of-sample datasets, with the last 12 observations reserved for validation as out-of-sample data. The regression model was fitted to the in-sample data, and predictions were made for the out-of-sample period using the significant regressors and the identified structural breaks.

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Performance Metrics

To assess the accuracy of the forecast, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE):

Table 5.14: Forecasting Performance Metrics for WTI Returns (Regression with Structural Breaks)

Metric	Value
MAE	0.0754
MSE	0.0088
RMSE	0.0940

These metrics provide insights into the forecast error and indicate the model's accuracy in predicting WTI returns with the inclusion of structural breaks.

Forecasting Results

The figure below presents the actual and predicted WTI returns over the final 12 months of the dataset:

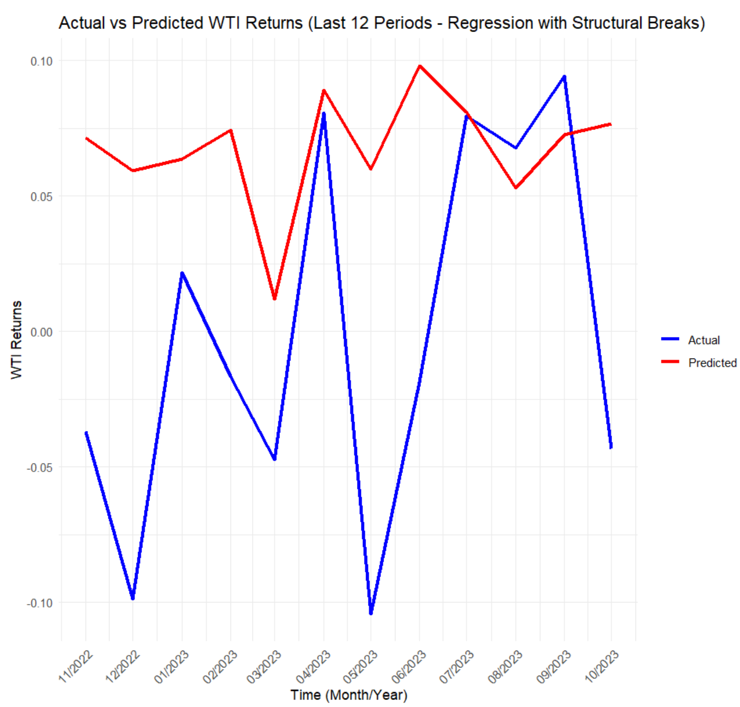

Figure 5.12: Actual vs Predicted WTI Returns (Last 12 Periods - Regression with Structural Breaks)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The model generally aligns with the observed data, though some deviations are noted, especially in volatile months, highlighting potential areas for further investigation.

5.6.4 Forecasting WTI Returns with ARMA(1,1) Model

To forecast the WTI returns for a 12-month period, we employed an ARMA(1,1) model with external regressors. The forecasting was conducted both with and without incorporating structural breaks. Structural breaks were identified using the Bai-Perron test, and the model with breaks included four dummy variables representing these breakpoints.

Model and Forecasting Process

The ARMA(1,1) model was fitted to the in-sample data, and predictions were made for the out-of-sample period. The model was evaluated with and without the inclusion of structural breaks in the regressors.

Performance Metrics

To assess the accuracy of the forecasts, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for both models:

Table 5.15: Forecasting Performance Metrics for WTI Returns (ARMA(1,1) Model)

Metric	**With Breaks**	**Without Breaks**
MAE	0.082	0.065
MSE	0.01	0.0058
RMSE	0.1	0.0765

These metrics indicate the average forecast error and provide a measure of the accuracy of our predictions with and without breaks.

Model Comparison

To determine if the inclusion of structural breaks improves the model, we computed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for both models:

Table 5.16: AIC and BIC Comparison for Models with and without Breaks

Metric	**With Breaks**	**Without Breaks**
AIC	-449.58	-432.06
BIC	-407.08	-403.73

UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The model with structural breaks yielded a lower AIC, suggesting a better fit compared to the model without breaks. An ANOVA test was also performed to confirm the improvement, indicating a significant difference with a p-value lower than 0.05.

Forecasting Results

The following figures present the actual and predicted WTI returns over the final 12 months for both models:

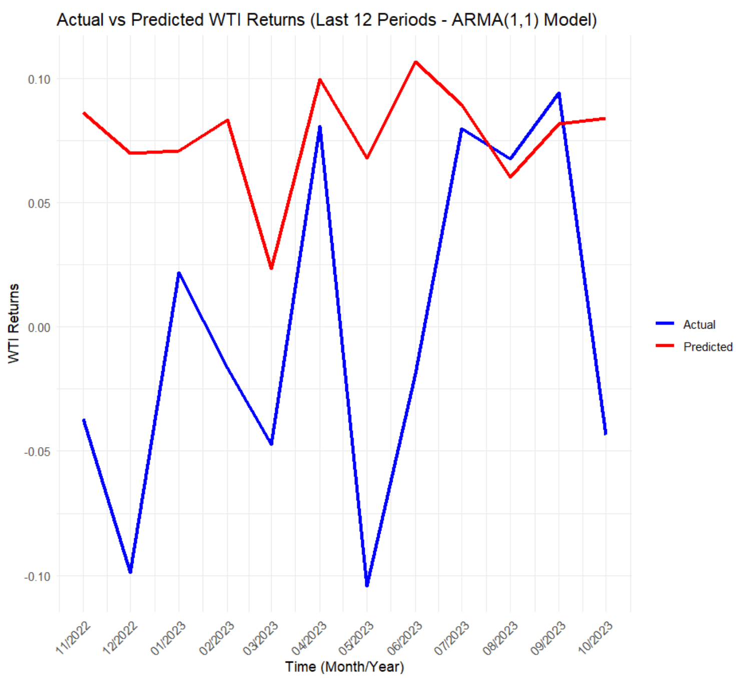

Figure 5.13: Actual vs Predicted WTI Returns (Last 12 Periods - ARMA(1,1) Model with Breaks)

In both plots, the blue line represents the actual WTI returns, while the red line represents the predicted values from the ARMA(1,1) models with and without structural breaks. The model with breaks provided a marginally better fit, as indicated by the performance metrics and information criteria, suggesting that the inclusion of structural breaks helps capture variations in the WTI returns.

Attribution 4.0 International https://doi.org/10.26219/heal.aueb.6444
http://creativecommons.org/licenses/by/4.0/

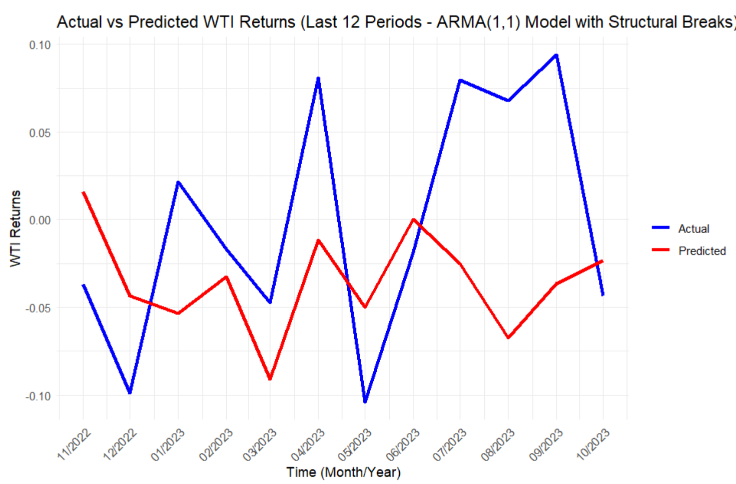

Figure 5.14: Actual vs Predicted WTI Returns (Last 12 Periods - ARMA(1,1) Model without Breaks)

5.6.5 Forecasting WTI Returns with Random Forest Model

To forecast WTI returns over a 12-month period, we employed a Random Forest model, utilizing the same variables as described earlier.

Model and Forecasting Process

The data was split into in-sample and out-of-sample datasets, where the last 12 observations were reserved as out-of-sample data for validation. A Random Forest model with 500 trees was trained on the in-sample data, and predictions were generated for the out-of-sample period.

Performance Metrics

To evaluate the forecast accuracy, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as follows:

Table 5.17: Forecasting Performance Metrics for WTI Returns (Random Forest Model)

Metric	**Value**
MAE	0.0625
MSE	0.0058
RMSE	0.0762

These metrics indicate the average forecast error and provide a measure of the accuracy of our predictions.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Forecasting Results

The figure below shows the actual and predicted WTI returns for the last 12 months of the dataset:

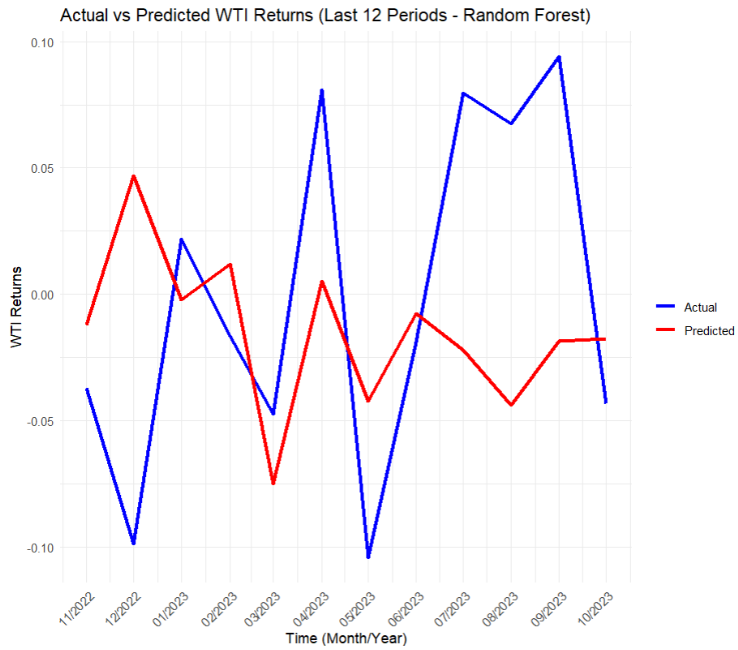

Figure 5.15: Actual vs Predicted WTI Returns (Last 12 Periods - Random Forest)

The model captures some fluctuations in the data, though differences between actual and predicted values in certain months suggest potential areas for improvement.

5.6.6 Forecasting WTI Returns with Support Vector Regression (SVR) Model

To forecast the WTI returns for a 12-month period, we used a Support Vector Regression (SVR) model. The statistically significant variables identified from stepwise regression were used as predictors.

Performance Metrics

To evaluate the forecast accuracy, we calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as follows:

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Table 5.18: Forecasting Performance Metrics for WTI Returns (SVR Model)

Metric	Value
MAE	0.0702
MSE	0.0069
RMSE	0.0832

Forecasting Results

The figure below shows the actual and predicted WTI returns for the last 12 months of the dataset:

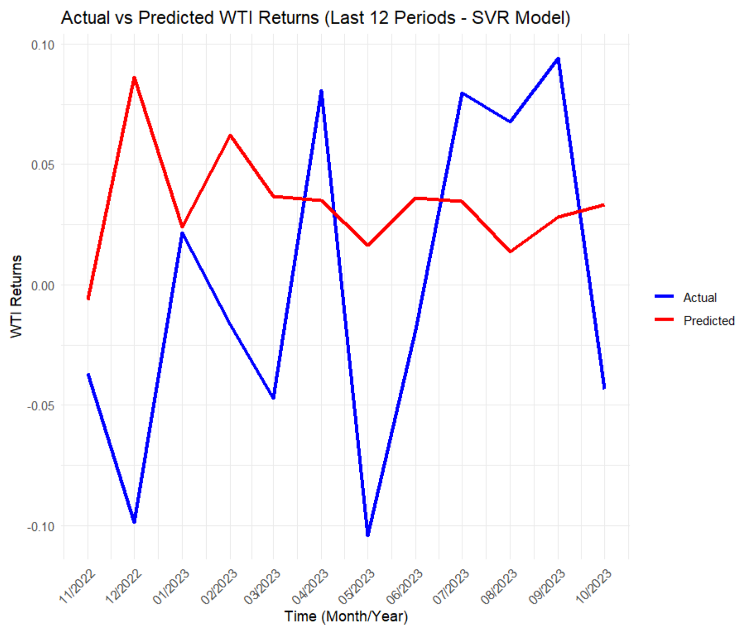

Figure 5.16: Actual vs Predicted WTI Returns (Last 12 Periods - SVR Model)

Some of the data's fluctuations are successfully reflected by the model. Differences between the actual and expected numbers in particular months, however, point to areas where the model needs more improvement.

5.6.7 Forecasting WTI Returns with Gradient Boosting Model

To forecast the WTI returns for a 12-month period, we utilized a Gradient Boosting Model (GBM). Statistically significant predictors, identified from stepwise regression, as previously.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Model and Forecasting Process

The Gradient Boosting Model was trained on the in-sample data with parameters set to 500 trees, a depth of 3, and a learning rate (shrinkage) of 0.01. Cross-validation was used to determine the optimal number of trees, resulting in 65 trees as the best iteration.

Performance Metrics

To evaluate the forecast accuracy, we computed the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as follows:

Table 5.19: Forecasting Performance Metrics for WTI Returns (Gradient Boosting Model)

Metric	Value
MAE	0.0583
MSE	0.0044
RMSE	0.0669

Forecasting Results

The figure below shows the actual versus predicted WTI returns over the last 12 months of the dataset:

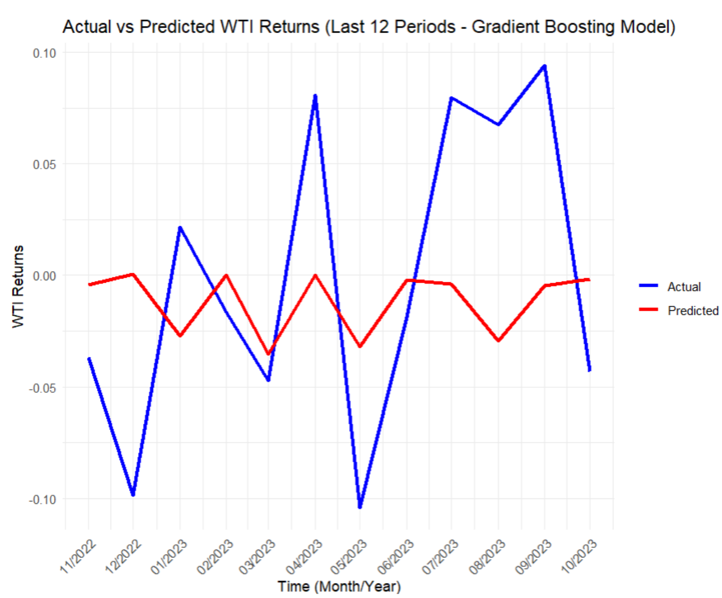

Figure 5.17: Actual vs Predicted WTI Returns (Last 12 Periods - Gradient Boosting Model)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

The model captures some patterns in the returns, though deviations suggest potential areas for further model improvement.

5.6.8 Summary of Forecasting Results for Brent Returns

To forecast the Brent returns a 12-month period, we use the full model without the variable Equity Market Volatility: Infectious Disease Tracker. The forecasting results for Brent crude oil returns using various models are summarized in the table below. The table presents the performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), for all models.

Table 5.20: Summary of Forecasting Performance Metrics for Brent Returns

Model	**MAE**	**MSE**	**RMSE**
Multiple Regression	0.0541	0.0040	0.0633
Rolling Window Regression	0.0554	0.0040	0.0635
Regression with Structural Breaks	0.0737	0.0085	0.0922
ARMA(1,1) (Without Breaks)	0.0544	0.0041	0.0643
ARMA(1,1) (With Breaks)	0.0884	0.0110	0.1050
Random Forest	0.0531	0.0047	0.0683
Support Vector Regression (SVR)	0.0506	0.0037	0.0610
Gradient Boosting Model (GBM)	0.0522	0.0038	0.0619

The table shows that the Support Vector Regression (SVR) and Gradient Boosting Model (GBM) achieved the best overall performance, with the lowest error metrics across MAE, MSE, and RMSE. Other models, such as Multiple Regression and Random Forest, also performed well, but with slightly higher error values.

The figure below shows the actual and predicted Brent returns for the last 12 months of the dataset:

The model captures some fluctuations in the data, though differences between actual and predicted values in specific months suggest potential areas for further model improvement. The figures for the other models are provided in the Appendix for reference.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

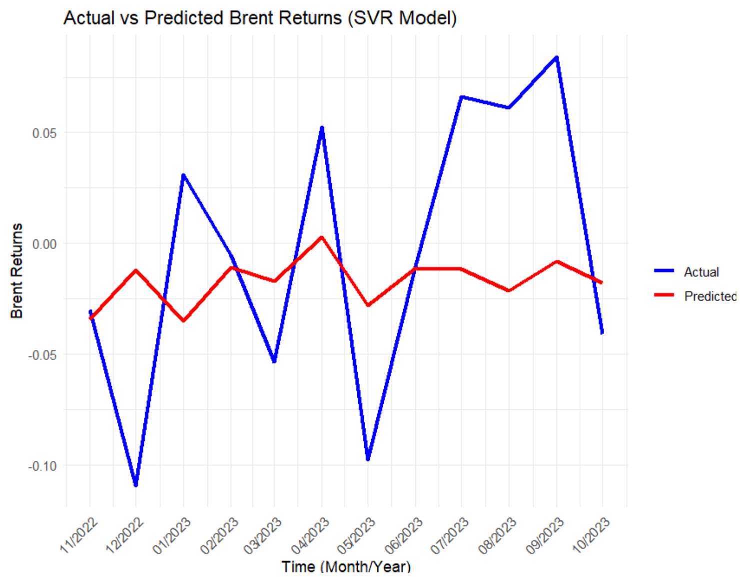

Figure 5.18: Actual vs Predicted Brent Returns (Last 12 Periods - SVR Model)

5.6.9 Summary of Forecasting Results for Dubai Returns

To forecast the Dubai returns a 12-month period, we use the full model without the variables: Equity Market Volatility: Infectious Disease Tracker and Federal Funds Effective Rate Various models were employed, including Multiple Regression, Rolling Window Regression, Regression with Structural Breaks, ARMA(1,1), Random Forest, Support Vector Regression (SVR), and Gradient Boosting Model (GBM). The performance metrics for each model are summarized in the table below.

Table 5.21: Summary of Forecasting Performance Metrics for Dubai Returns

Model	**MAE**	**MSE**	**RMSE**
Multiple Regression	0.0622	0.0047	0.0686
Rolling Window Regression	0.0634	0.0048	0.0693
Regression with Structural Breaks	0.0593	0.0060	0.0776
ARMA(1,1) (Without Breaks)	0.0629	0.0050	0.0706
ARMA(1,1) (With Breaks)	0.0757	0.0093	0.0964
Random Forest	0.0588	0.0049	0.0703
Support Vector Regression (SVR)	0.0560	0.0041	0.0640
Gradient Boosting Model (GBM)	0.0573	0.0044	0.0662

The table indicates that the Support Vector Regression (SVR) model achieved the best overall performance, with the lowest MAE, MSE, and RMSE values. The Gradient Boosting Model (GBM) and Multiple Regression also performed

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

well, but with slightly higher error metrics.

The figure below shows the actual and predicted Dubai returns over the final 12 months of the dataset using the SVR model:

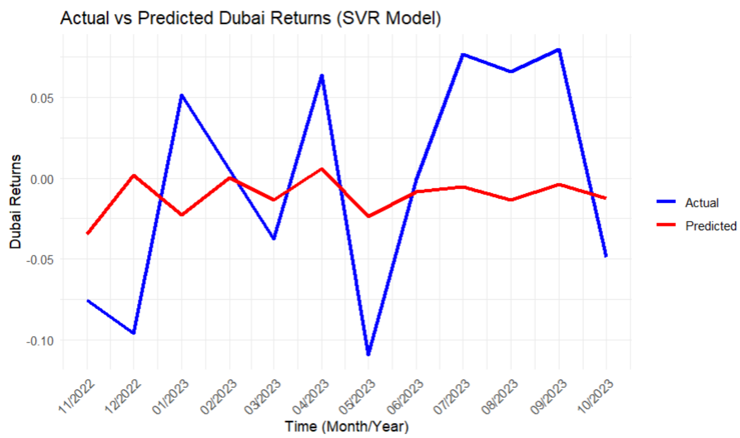

Figure 5.19: Actual vs Predicted Dubai Returns (Last 12 Periods - SVR Model)

Although certain data fluctuations are captured by the model, differences between actual and expected values in particular months point to possible areas for future model development. For reference, the Appendix contains the figures for the various models.

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Chapter 6

Conclusions and Further Research

6.1 Conclusions

The dynamics of crude oil price returns have been examined in this thesis using advanced econometric models and machine learning methodologies. This study provides insight into price volatility, structural breaks, and the interaction between market dynamics and external shocks by analyzing three important benchmarks: WTI, Brent, and Dubai crude.

The combination of modern machine learning techniques like random forests and gradient boosting with traditional approaches like ARMA-GARCH has produced encouraging outcomes. While econometric approaches provided interpretability and theoretical foundation, machine learning models showed their ability to capture non-linear correlations and adjust to the high-frequency nature of oil market data. This combination emphasizes how crucial it is to integrate several approaches in order to attain strong prediction performance.

The importance of structural breaks, which represent geopolitical and economic crises, in influencing the dynamics of oil prices is one of the main conclusions. Both econometric and machine learning estimates were more accurate when such pauses were included. Furthermore, outside variables like policy decisions and the state of the world economy have become important predictors of the movement of the oil price.

Despite these successes, difficulties still exist. Although machine learning methods have improved predicted accuracy, their complexity makes them difficult to understand and apply in situations involving policymaking. Similar to this, despite their strength, econometric models can find it difficult to adjust to the quickly changing environment of the oil market.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

6.2 Further Research

This study provides a number of opportunities for more research. First, adding other data sources, like real-time news, could improve the understanding of changes in the oil market. By incorporating such data into current models, predicting accuracy may be increased and hidden patterns may be revealed.

Second, there is potential in investigating hybrid models that further combine econometric and machine learning methods. For example, GARCH-family models combined with deep learning architectures may be able to capture volatility clustering more accurately while preserving theoretical robustness.

Finally, it is crucial to understand how traditional and developed energy markets interact as the world's energy industry shifts to renewable sources. In order to provide a comprehensive understanding of the dynamics of the energy market, future research might concentrate on integrating environmental indexes and renewable energy measures into forecasting models.

To sum up, this thesis has established a framework for integrating conventional and modern techniques to tackle the complex nature of crude oil markets. Navigating the changing issues of energy economics will need expanding on these results through additional interdisciplinary study.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International https://doi.org/10.26219/heal.aueb.6444
http://creativecommons.org/licenses/by/4.0/

Bibliography

Bai, J., and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica, 66*(1), 47-78. https://doi.org/10.2307/2998540

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307-327.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chai, T., and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427-431. https://doi.org/10.1080/01621459.1979.10482531

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4), 987-1007. https://doi.org/10.2307/1912773

Ewing, B. T., and Malik, F. (2017). Volatility transmission and correlation in oil and natural gas markets. *Energy Economics, 64*, 93-102. https://doi.org/10.1016/j.eneco.2017.02.017

Ferderer, J. P. (1996). Oil price volatility and the macroeconomy. *Journal of Macroeconomics, 18*(1), 1-26. https://doi.org/10.1016/S0164-0704(96)80001-2

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance, 48*(5), 1779-1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Attribution 4.0 International https://doi.org/10.26219/heal.aueb.6444
http://creativecommons.org/licenses/by/4.0/

Guo, H., and Kliesen, K. L. (2005). Oil price volatility and U.S. macroeconomic activity. *Federal Reserve Bank of St. Louis Review, 87*(6), 669-683.

Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Koenker, R., and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143-156. https://doi.org/10.1257/jep.15.4.143

Lahmiri, S. (2017). A comparison of PNN and SVM for stock market index prediction: Empirical study on the Casablanca Stock Exchange. *Physica A: Statistical Mechanics and its Applications*, 471, 205-212. https://doi.org/10.1016/j.physa.2016.12.018

Li, J., and Ge, H. (2013). Support vector regression for crude oil price forecasting. *Energy Economics*, 40, 485-494. https://doi.org/10.1016/j.eneco.2013.07.010

Liu, L., Geng, Q., Zhang, Y., and Wang, Y. (2022). Investors' perspective on forecasting crude oil return volatility: Where do we stand today? *Journal of Management Science and Engineering, 7*(4), 423-438. https://doi.org/10.1016/j.jmse.2022.06.001

Malik, F., and Ewing, B. T. (2009). Volatility transmission between oil prices and equity sector returns. *International Review of Financial Analysis, 18*(3), 95-100. https://doi.org/10.1016/j.irfa.2009.03.003

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica, 59*(2), 347-370. https://doi.org/10.2307/2938260

OECD. (2022). Competition in Energy Markets. OECD Competition Policy Roundtable.

Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Economics, 21*(5), 449-469. https://doi.org/10.1016/S0140-9883(99)00020-1

Smola, A. J., Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0

Taylor, S. J. (1986). *Modelling Financial Time Series.* Wiley.

Yu, L., Zhao, Y., Tang, L., and Zhang, Y. (2016). A least squares support vector machine approach for crude oil price forecasting: A case study. *Energy*, 94, 540-549. https://doi.org/10.1016/j.energy.2015.11.016

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control, 18*(5), 931-955. https://doi.org/10.1016/0165-1889(94)90039-6

Zhang, X., Wang, C., and Ma, Y. (2017). Machine learning for financial time series forecasting: A survey. *IEEE Access, 7*, 12367-12384. https://doi.org/10.1109/ACCESS.2017.2764659

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

Appendix

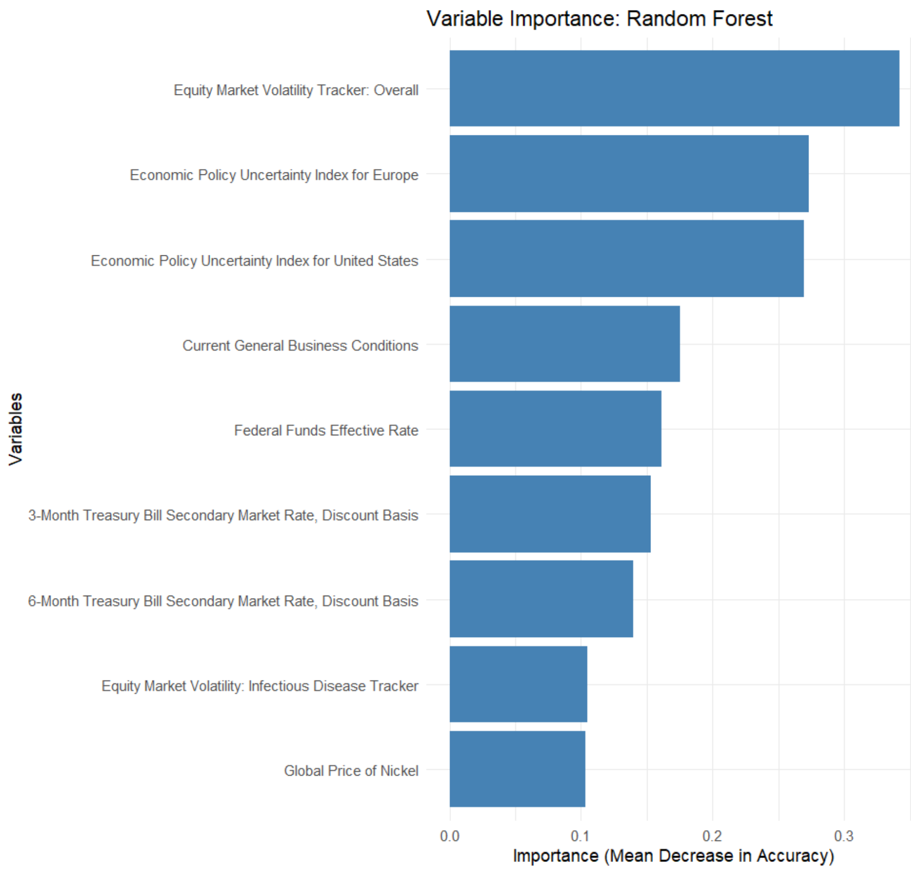

Figure 6.1: Variable Importance for Brent Returns (Random Forest Model)

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

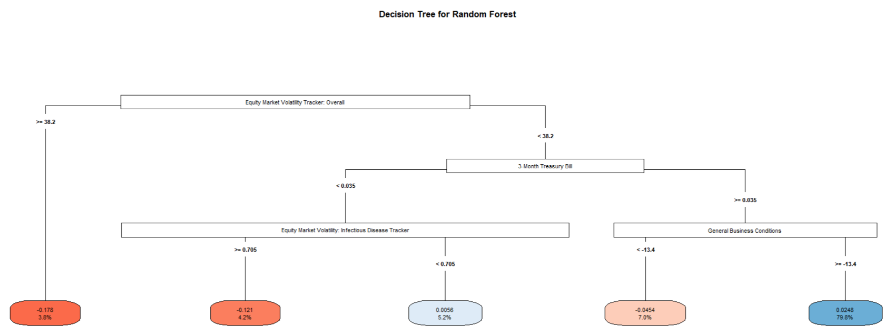

Figure 6.2: Decision Tree for Brent Returns (Random Forest Model)

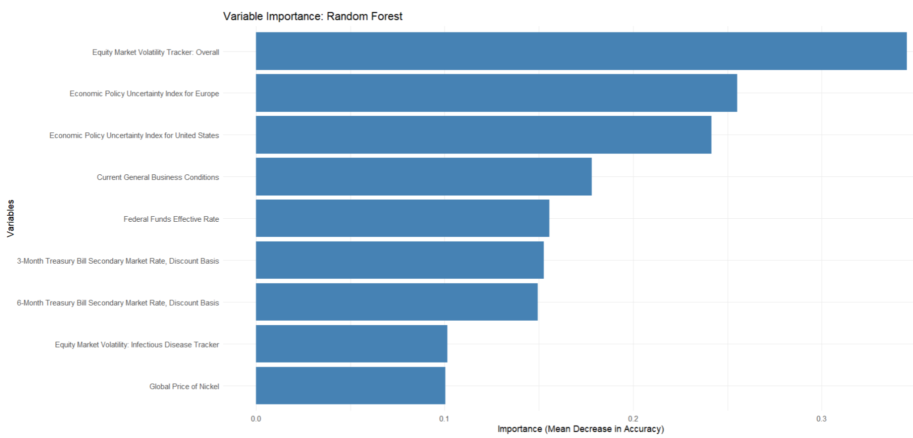

Figure 6.3: Variable Importance for Dubai Returns (Random Forest Model)

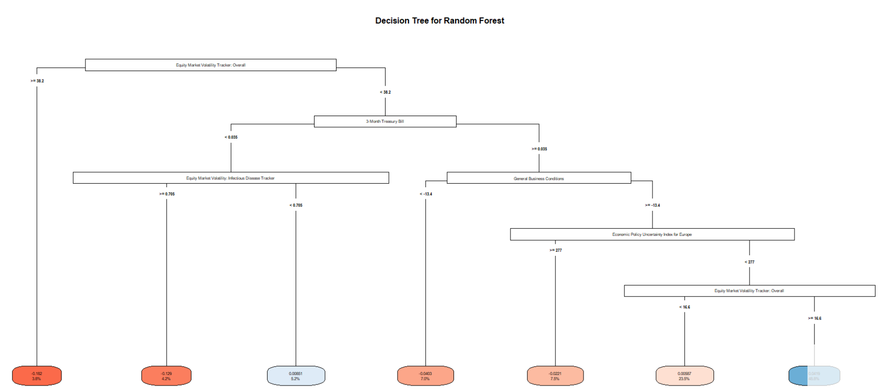

Figure 6.4: Decision Tree for Dubai Returns (Random Forest Model)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

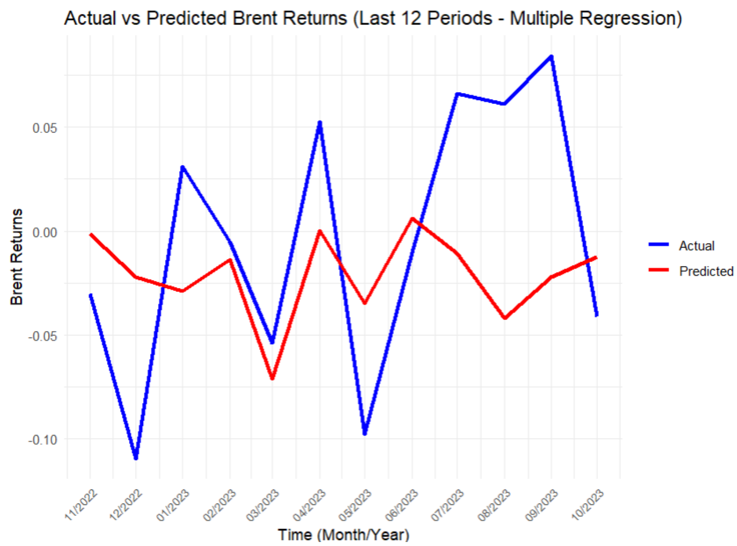

Figure 6.5: Actual vs Predicted Brent Returns (Last 12 Periods - Multiple Regression)

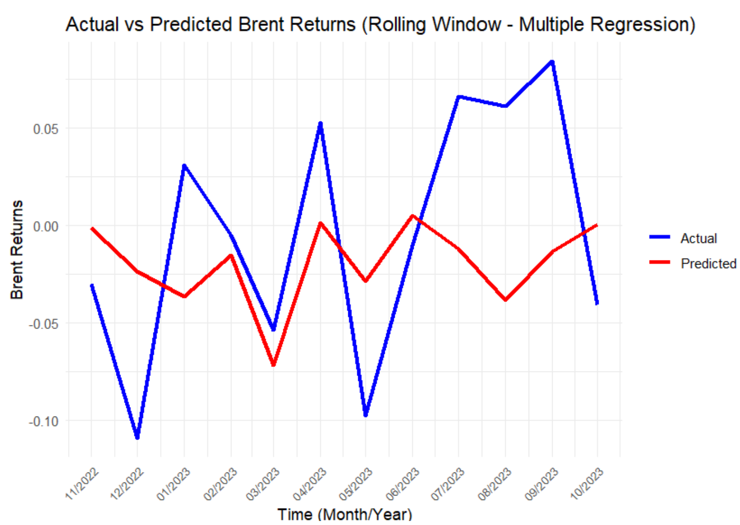

Figure 6.6: Actual vs Predicted Brent Returns (Rolling Window - Multiple Regression)

Attribution 4.0 International
https://doi.org/10.26219/heal.aueb.6444
http://creativecommons.org/licenses/by/4.0/

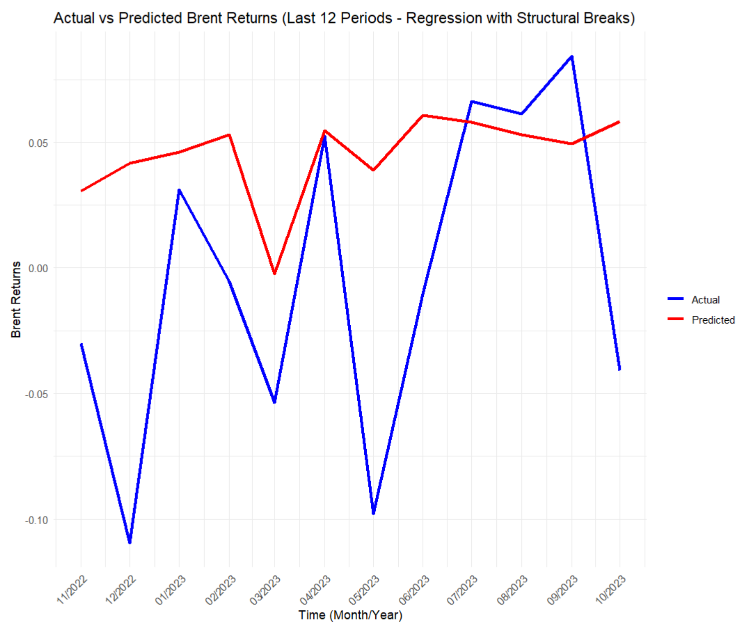

Figure 6.7: Actual vs Predicted Brent Returns (Last 12 Periods - Regression with Structural Breaks)

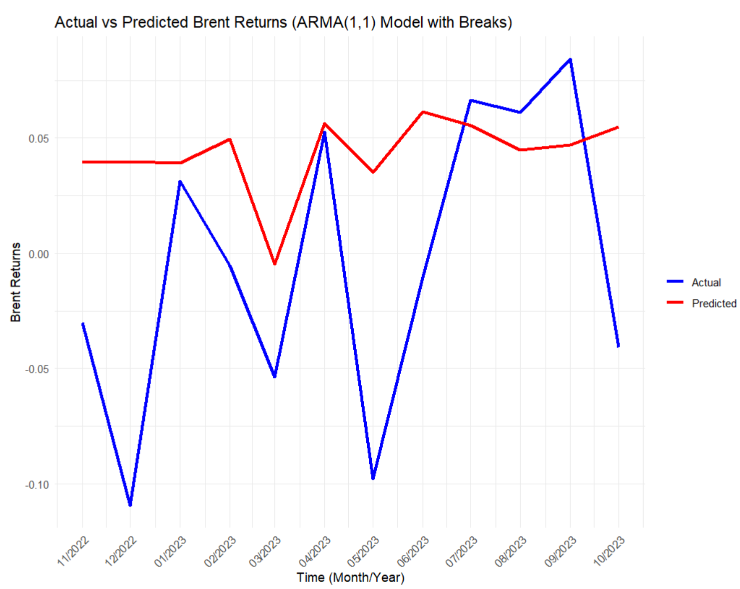

Figure 6.8: Actual vs Predicted Brent Returns (Last 12 Periods - ARMA(1,1) Model with Breaks)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

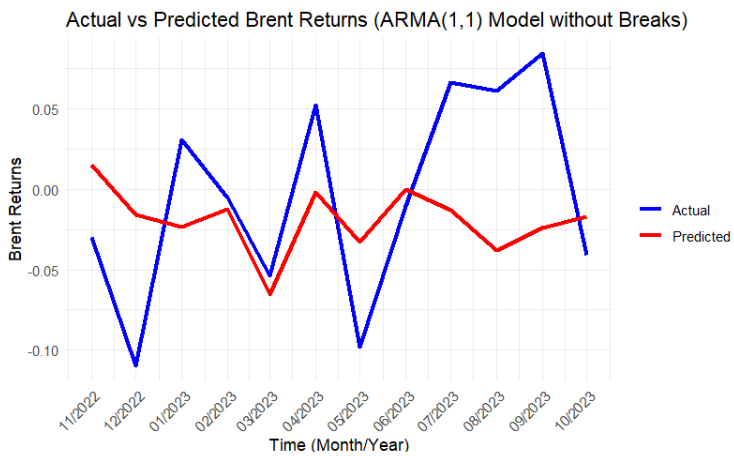

Figure 6.9: Actual vs Predicted Brent Returns (Last 12 Periods - ARMA(1,1) Model without Breaks)

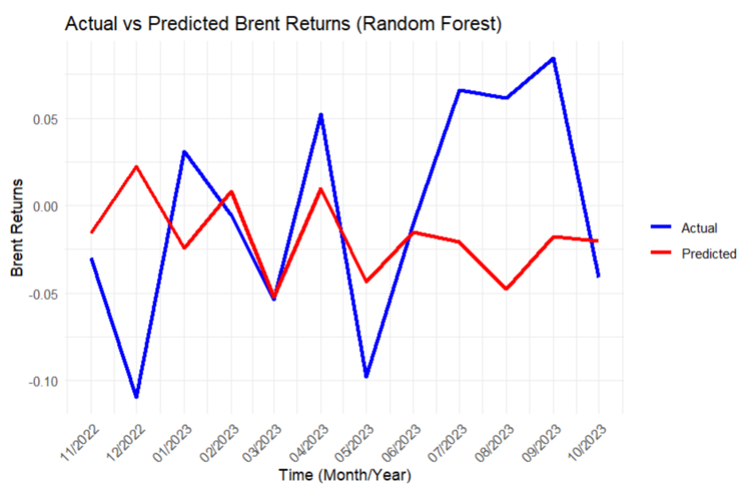

Figure 6.10: Actual vs Predicted Brent Returns (Last 12 Periods - Random Forest)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

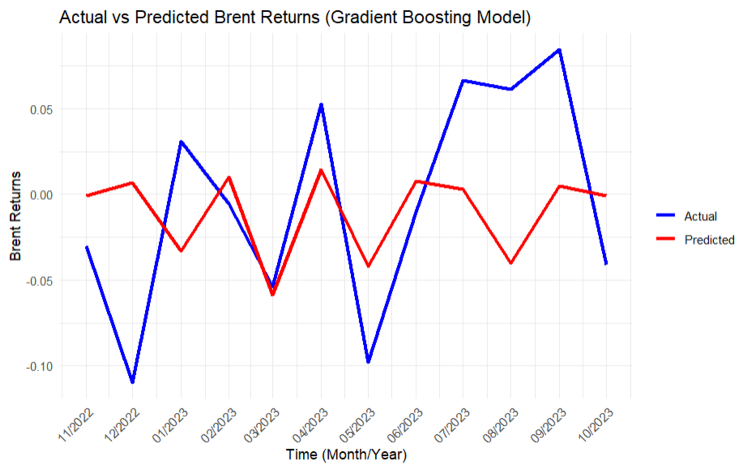

Figure 6.11: Actual vs Predicted Brent Returns (Last 12 Periods - Gradient Boosting Model)

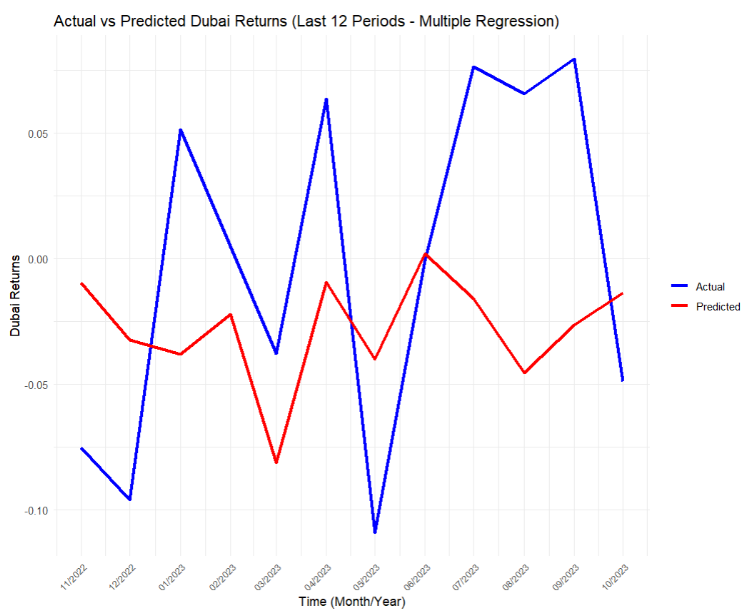

Figure 6.12: Actual vs Predicted Dubai Returns (Last 12 Periods - Multiple Regression)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

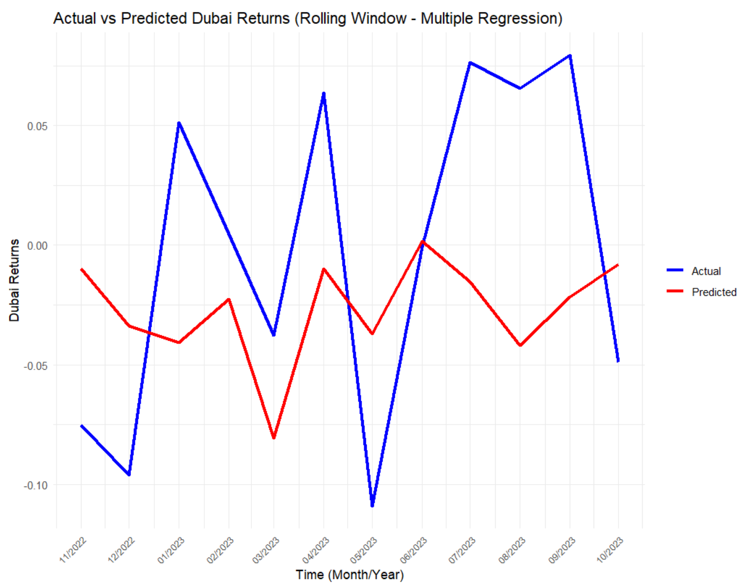

Figure 6.13: Actual vs Predicted Dubai Returns (Rolling Window - Multiple Regression)

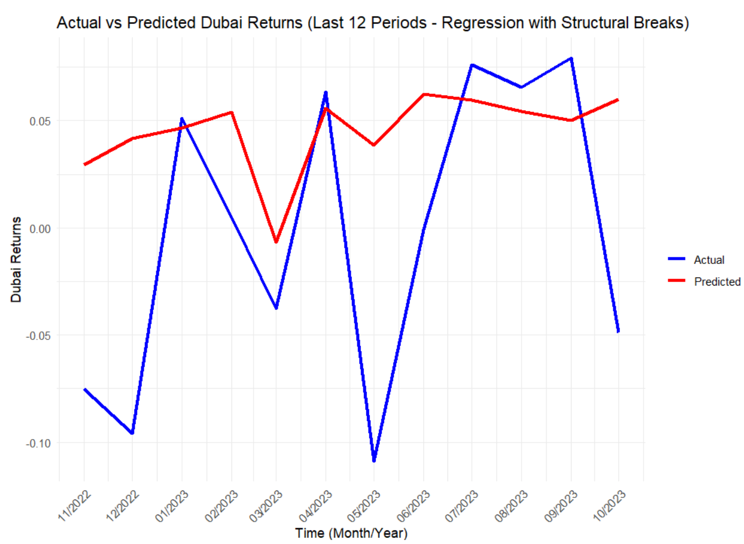

Figure 6.14: Actual vs Predicted Dubai Returns (Last 12 Periods - Regression with Structural Breaks)

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.6444

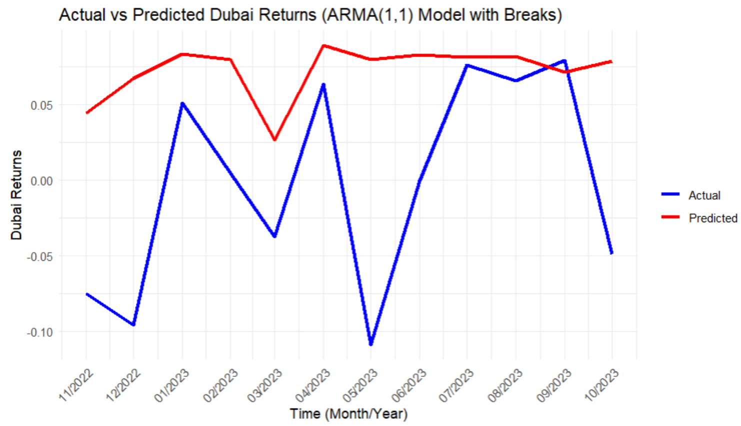

Figure 6.15: Actual vs Predicted Dubai Returns (Last 12 Periods - ARMA(1,1) Model with Breaks)

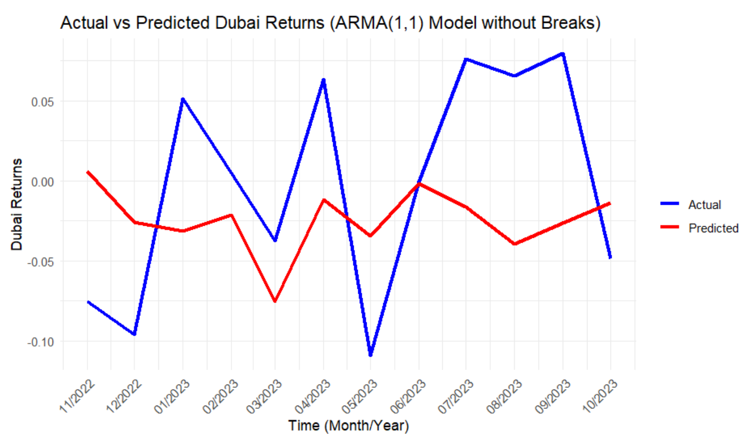

Figure 6.16: Actual vs Predicted Dubai Returns (Last 12 Periods - ARMA(1,1) Model without Breaks)

Attribution 4.0 International
http://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.6444

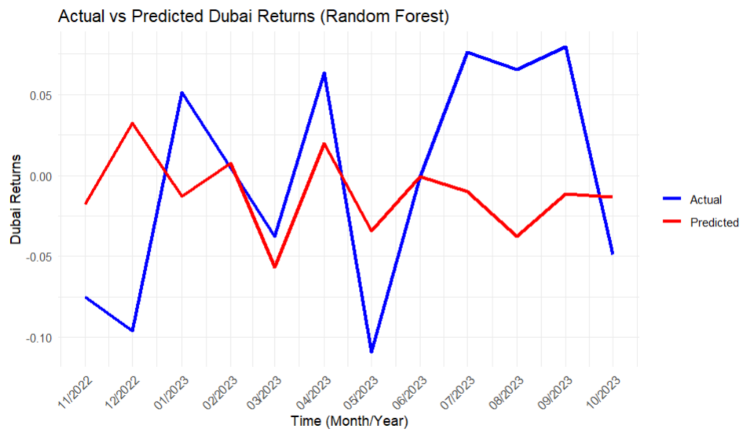

Figure 6.17: Actual vs Predicted Dubai Returns (Last 12 Periods - Random Forest)

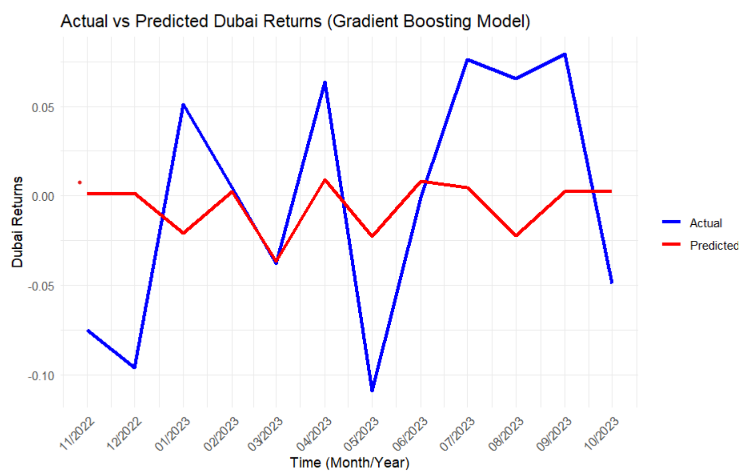

Figure 6.18: Actual vs Predicted Dubai Returns (Last 12 Periods - Gradient Boosting Model)