AI for Finance: Final Project
Forecasting WTI Crude Oil Spot Prices

Giulio Mattolin 224072

Summary

In this work, the problem of forecasting the spot prices of West Texas Intermediate crude oil was addressed by considering and comparing two different approaches, performing both an econometric and a neural network analysis. Historic data about multiple variables over a period of 23 years were collected and preprocessed in order to organize and clean them. In the econometric analysis, GARCH models were used to extract conditional volatilities and linear models were used to forecast WTI spot prices. On the other hand, in the neural network analysis, a LSTM-based model was used to take into account the sequentiality of the data. Looking at the obtained results, it is possible to conclude that the neural network model outperformed the econometric models in forecasting WTI spot prices. Indeed, the neural network predictions are more accurate and able to capture the correct trend and variability of the prices. Code and data are available at the following link: https://colab.research.google.com/drive/1Hua9dAU5xUUR13L5s1Sx7_I4IQ2iq0MH.

1 Introduction

Crude oil plays a significant role in the global economy, since approximately 39% of the world's energy consumption comes from it. The price of oil is a major factor in the overall health of the energy sector and is one of the most heavily traded commodities as it is influenced by almost every global, macro event. Fluctuations in the price of oil can have a significant impact on the economies of countries that export or import it, so being able to predict oil prices accurately can assist policymakers in making informed decisions about energy resources and effective policies. Crude oil price prediction has been a challenging problem in forecasting research because its prices are affected by many factors. Except for the fundamental market factors, such as supply, demand and inventory, oil price fluctuation is strongly influenced by economic development, conflicts, wars and breaking news.

Brent Crude and West Texas Intermediate (WTI) are the world's two major oil markets. Brent Crude is extracted from oil fields in the North Sea, between the Shetland Islands and Norway, and is the benchmark used for the light oil market in Europe, Africa, and the Middle East. On the other hand, West Texas Intermediate is sourced from U.S. oil fields, primarily from inland Texas, and it is the benchmark for the U.S. light oil market.

In the following work, the price spot of West Texas Intermediate is taken into consideration with the aim of forecasting future changes in its value. In particular, data regarding stocks of companies working in the oil industry, data related to the oil economy and few macroeconomic indicators were collected. The data were then preprocessed, in order to clean it and obtain a common format, and later an exploratory data analysis was carried out in order to understand the relationships among the different variables and their distributions. After that, through the use of econometrics methods and neural networks, the data was leveraged to study the movements of WTI spot prices.

The remaining of the paper is organized as follows. Section 2 illustrates the data, presents the preprocessing steps performed along with a descriptive analysis of the data. Section 3 introduces the models used in the work. Section 4 presents the implementation details and the results obtained from the econometric analysis. Section 5 presents the implementation details and the results obtained in the neural network analysis. Lastly, section 6 draws conclusions about the overall project and the achieved results.

2 Data description

The data used to analyze WTI spot prices range from January 1999 to April 2022 and has a monthly frequency. In particular, 27 different variables, including WTI spot prices variable, were collected from 3 sources and are illustrated in Table 1 divided into 4 main categories.

The first typology regards stock market data which can be useful to track how the oil market is moving. Therefore, stock data about 11 of the biggest oil companies in the world were obtained from Refinitiv. In order to have a varied and representative sample of data, companies from different countries were selected but the majority of them are from the US since WTI crude oil is extracted there. To note that the start of the data is January 1999 and not an older date since many modern big oil companies do not have a long history and thus their stock market data availability is limited.

The second typology regards stock market indices data which provide a broad view of how the markets and economy of certain regions or sectors are performing. In particular, data from 2 indices measuring the performance of US and EU markets were collected from Refinitiv, namely S&P 500 and STOXX Europe 600.

The third typology regards oil economy data, such as prices of WTI crude oil and Brent crude oil, WTI future contracts, the import/export of WTI and the production and consumption of it. These data were obtained from the U.S. Energy Information Administration (EIA) website and they are important since they are directly associated with WTI crude oil.

Lastly, the fourth typology regards macroeconomic data such as inflation, recession and events that have a significant impact on the global economy, like pandemics and wars. In particular, data concerning inflation and recession were obtained from the Federal Reserve Bank of St. Louis website, while data about COVID-19 and the Russia-Ukraine War were added as dummy variables based on the start dates of these events.

2.1 Data cleaning

Once the data were collected, mainly in .csv file format, some preprocessing steps were performed in order to arrange all the data in the same format and join the different information into a single table.

Primary type	Secondary type	Variables	Source
Stocks	US	Exxon Mobil Corporation (XOM)	Refinitiv
		Chevron Corporation (CVX)	Refinitiv
		Marathon Petroleum Corporation (MRO)	Refinitiv
		Valero Energy Corporation (VLO)	Refinitiv
		ConocoPhillips Company (COP)	Refinitiv
	UK	Shell plc (SHEL.AS)	Refinitiv
		BP plc (BP.L)	Refinitiv
	IN	Indian Oil Corporation Limited (IOC.NS)	Refinitiv
		Reliance Petroleum Limited (RELI.NS)	Refinitiv
	FR	TotalEnergies SE (TOT.TO)	Refinitiv
	IT	Eni S.p.A (ENI.MI)	Refinitiv
Index	US	S&P500 (.SP500)	Refinitiv
	EU	STOXX 600 (.STOXX)	Refinitiv
Oil economy	Prices	WTI Spot Price Dollars per Barrel	EIA
		Brent Spot Price Dollars per Barrel	EIA
		Cushing OK Future Contract 1 Dollars per Barrel	EIA
		Cushing OK Future Contract 2 Dollars per Barrel	EIA
		Cushing OK Future Contract 3 Dollars per Barrel	EIA
		Cushing OK Future Contract 4 Dollars per Barrel	EIA
	Imports/Exports	U.S. Exports of Crude Oil Thousand Barrels	EIA
		U.S. Imports of Crude Oil Thousand Barrels	EIA
	Production	U.S. Field Production of Crude Oil Thousand Barrels	EIA
	Consumption	U.S. Product Supplied of Crude Oil and Petroleum Products Thousand Barrels	EIA
Macroeconomic	Inflation	Sticky Price Consumer Price Index	St. Louis Fed
	Recession	U.S. Recession	St. Louis Fed
	Pandemics	COVID-19	/
	Wars	Russia–Ukraine War	/

Table 1: Collected data organized by typology.

In particular, using the library *pandas* of Python, all variables' data were loaded into individual DataFrames and their Date columns were changed to have the same format, i.e. YEAR-MONTH. After that, all data were merged together into a single DataFrame keeping as reference the Date column. In order to have all data starting from a common date, since some variables had data availability also before January 1999, the DataFrame was trimmed to start from January 1999, thus removing data prior to this date. Lastly, once a uniform table containing all variables with the same format was obtained, the data was converted to stationary by calculating the percentage change between the current and the prior month, for all months.

In order to fairly assess the performance of the forecasting model, the data was then split into two datasets, namely train set and test set. In this way, models are tested on unseen data and it is possible to measure how effectively they can generalize. Specifically, the first 48 months, i.e. 4 years, of data were considered as test set, while the rest as train set. It was decided to take the first 4 years and not the last 4 years as test set because both COVID-19 and the Russia-Ukraine War broke out in the last two years, and thus the model would not have been trained on these events, ending up with poor and unfair performance. It is possible to visualize the data split in Figure 1.

2.2 Descriptive statistics

Before fitting the models and forecasting the prices of WTI crude oil, it is important to analyze and investigate the characteristics of the dataset. In this way, it is possible to discover patterns, spot anomalies and check assumptions. To this end, the distributions of

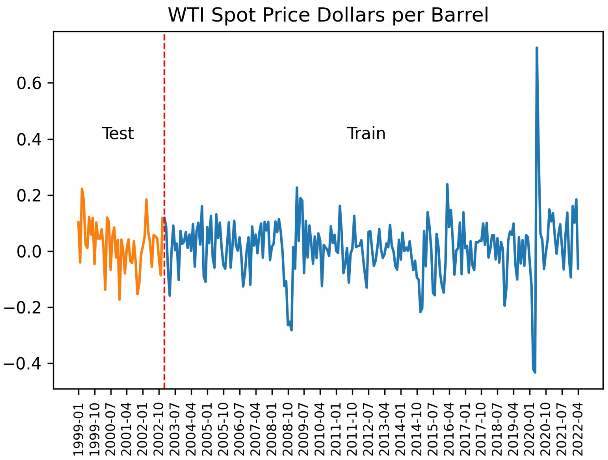

Figure 1: Visualization of train and test split.

variables, their correlations and possible anomalies were explored.

In Figure 2, the distributions of the preprocessed variables are visualized through box-plots. To note that the dummy variables related to COVID-19 and Russia-Ukraine War were not considered for this plot since they have only 0s and 1s as values. As it is possible to notice, most variables have a distribution between [-1, 1], except for the export of WTI variable which has values greater than 1 and also an extreme outlier with value greater than 12. To investigate this anomaly, plots of the variable about the export of WTI crude oil before and after the stationarity preprocessing are visualized in Figure 3 and 4. By looking at these plots, it can be concluded that these large values are the consequence of a period of very low exportation of WTI crude oil up to 2015. This is the consequence of the crude oil export ban in US which lasted from 1977 to 2015 under the Energy Policy and Conservation Act [4]. Even if the changes in that period appear small, considering the entire Figure 3, it happened for example that in April 2001 there were 145 thousand barrels of crude oil exported while on May 2001 there were 1981 thousand barrels of crude oil exported, causing a percentage change of 12.6.

Additionally, the correlation among variables was measured and visualized in Figure 5 through an heatmap. As it is possible to notice, some variables have a high positive correlation and none have a high negative correlation. In particular, there is a notable correlation among the variables related to the oil companies stocks and a high correlation among the Brent crude oil, WTI crude oil and its future contracts. To better visualize relevant correlations, the ones with an absolute value lower than 0.5 are hidden in Figure 6. It is important to observe how the spot price of WTI crude oil is highly correlated with the one of Brent crude oil, with a coefficient of 0.95. Moreover, there is perfect multicollinearity between the spot price of WTI and its 1-month future contract, which is also highly correlated to the other future contracts.

Figure 2: Boxplot visualization of variables' distribution.

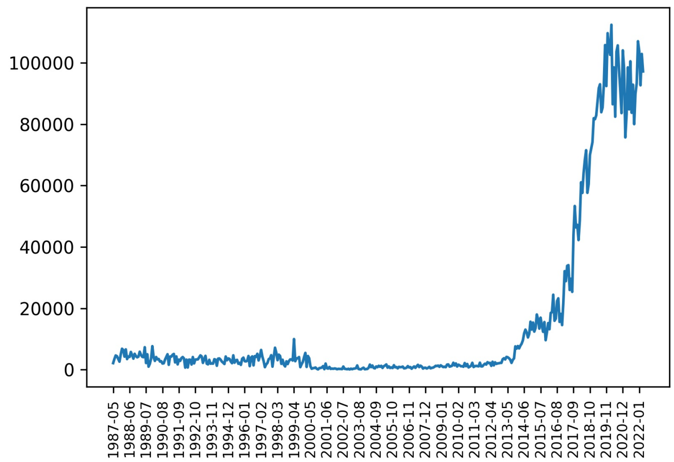

Figure 3: Visualization of export variable before stationarity preprocessing.

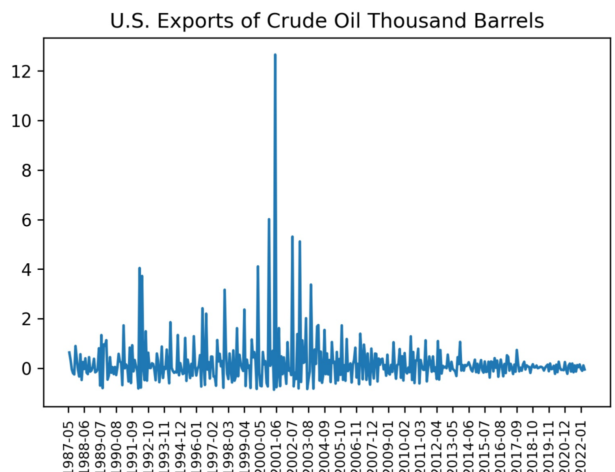

Figure 4: Visualization of export variable after stationarity preprocessing.

Figure 5: Heatmap visualization of variables' correlation.

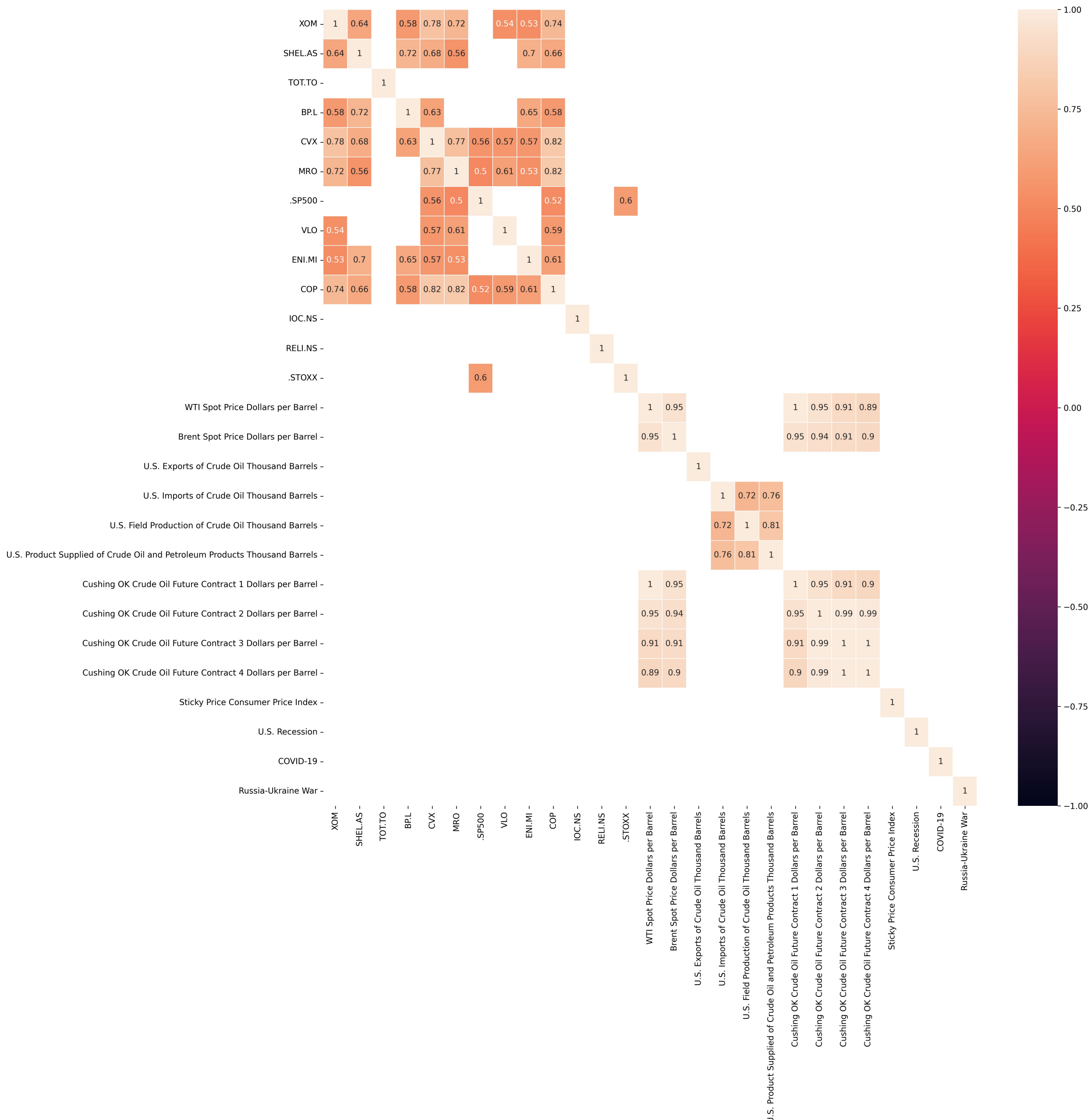

Figure 6: Heatmap visualization of variables' correlation over 0.5.

3 Forecasting models

The objective of this project is to forecast the spot price of WTI crude oil by leveraging the previously illustrated explanatory variables. To this end, both models belonging to econometric and deep learning are used, and their performance is measured and compared.

In the econometric analysis, GARCH models are used to account for the heteroscedasticity of explanatory variables and estimate their volatility. Explanatory variables, together with the conditional volatility of those having ARCH effect, are used then to forecast WTI crude oil prices employing linear models, both static and dynamic. The models used in the econometric analysis were built taking advantage of the library *statsmodels*.

In the neural network analysis, a specific deep learning model called long short-term memory (LSTM) [1] is used to forecast WTI crude oil prices. The LSTM architecture is particularly efficient at modeling time series since it takes into account the sequentiality of data and is capable of learning long-term dependencies. The model used in the neural network analysis was built taking advantage of the framework *PyTorch*.

4 Econometric analysis

The descriptive analysis showed that some of the collected variables are highly correlated. When dealing with linear models, multicollinearity is a problem since it violates the assumption of independence between explanatory variables. To overcome this, in the following econometric analysis the 3 variables *Cushing OK Future Contract 1 Dollars per Barrel*, *Cushing OK Future Contract 2 Dollars per Barrel* and *Cushing OK Future Contract 3 Dollars per Barrel* were excluded due to their high correlation with *Brent Spot Price Dollars per Barrel* and *Cushing OK Crude Oil Future Contract 4 Dollars per Barrel*, ending up with 23 independent variables in the dataset that can be used to forecast WTI spot prices.

4.1 GARCH models

Following the removal of problematic variables, GARCH models were used to analyze the conditional volatility of the various variables and include useful additional data in the dataset. A GARCH model for each variable was fitted and the p-value associated to the coefficient α_1 was checked. In particular, if the p-value of α_1 is lower than 0.05, it is statistically significant and the presence of ARCH effect is confirmed. The conditional volatility of all variables with α_1 statistically significant was then extracted and added to the dataset as an explanatory variable. Of the 23 independent variables, 10 had ARCH effect. In Table 2 it is possible to inspect the variables whose conditional volatility was added to the dataset and the p-values of their α_1 coefficient.

Variables	α_1 p-value
.SP500	0.0011
BP.L	0.0447
COP	0.0249
Cushing OK Crude Oil Future Contract 4 Dollars per Barrel	0.0287
MRO	0.0087
SHEL.AS	0.0369
Sticky Price Consumer Price Index	0.0270
U.S. Product Supplied of Crude Oil and Petroleum Products Thousand Barrels	0.0003
U.S. Recession	1.9e-19
XOM	0.0316

Table 2: Explanatory variables with ARCH effect.

4.2 Static models

With these data, namely the 23 independent variables and the conditional volatility of 10 of those, 12 linear models with each an additional lag in the explanatory variables were fitted on the train set. On each run, the statistically significant variables, thus with a p-value lower than 0.05, were stored with their frequency in order to track robust predictors that may improve the performance of the model. Moreover, the accuracy of the models was measured on the test set by calculating the mean squared error (MSE) between the predicted values and the ground truth. The results are reported in Table 3.

Lags	1	2	3	4	5	6	7	8	9	10	11	12
MSE	0.0112	0.0093	0.0186	0.0243	0.0155	0.0122	0.0083	0.0127	0.0100	0.0089	0.0096	0.0121

Table 3: Performance of static linear models on test set.

As it is possible to notice, the model with the best performance on the test set is the one with 7 lags achieving a MSE equal to 0.0083. Moreover, by inspecting the p-values of the predictors, this model found statistically significant only 2 variables, namely the conditional volatility of *COP* and *MRO*.

4.3 Dynamic models

The same process was executed using dynamic linear models, which consider as an explanatory variable also the response variable with lag, in order to take into account autoregression. In addition to the response variable with lag, it was also taken into account its conditional volatility extracted by a GARCH model, since its α_1 resulted statistically significant. Therefore, other 12 linear models with each an additional lag in the explanatory variables were fitted on the train set composed of 35 variables and the results are reported in Table 4.

Lags	1	2	3	4	5	6	7	8	9	10	11	12
MSE	0.0116	0.0092	0.0195	0.0257	0.0159	0.0122	0.0084	0.0129	0.0097	0.0087	0.0097	0.0128

Table 4: Performance of dynamic linear models on test set.

Once again, the model with the best performance on the test set is the one with 7 lags achieving a MSE equal to 0.0084, and the statistically significant predictors are the conditional volatility of *COP* and *MRO*. However, as it can be seen, by introducing autoregression the models did not improve their performance, probably due to high correlation between the response variable and other other variables, such as *Brent Spot Price Dollars Per Barrel*.

4.4 Robust models

By inspecting the robust variables tracked during these 24 runs, 12 with static models and 12 with dynamic models, and their frequency of being statistically significant illustrated in Table 5, it is possible to notice how few predictors resulted statistically significant in many runs and, interestingly, none of them were initially strongly correlated with the response variable. The most robust variables are: *U.S. Recession* and its conditional volatility, *CVX*

Robust predictors	Frequency
U.S. Recession	6
U.S. Recession CondVol	6
CVX	6
COVID-19	4
XOM	2
XOM CondVol	2
IOC.NS	2
ENI.MI	2
COP CondVol	2
SHEL.AS	2
SHEL.AS CondVol	2
MRO	2
MRO CondVol	2
RELI.NS	2
.SP500	2
.SP500 CondVol	2
Sticky Price Consumer Price Index	2
TOT.TO	2
U.S. Product Supplied of Crude Oil and Petroleum Products Thousand Barrels	1

Table 5: Robust predictors and their frequency of being statistically significant.

and *COVID-19*. Since other predictors were only statistically significant in 1 or 2 runs out of 24, it was decided to consider these first four predictors as robust and use them as explanatory variables to fit another 12 linear models with each an additional lag. The results are reported in Table 6.

Lags	1	2	3	4	5	6	7	8	9	10	11	12
MSE	0.0102	0.0098	0.0099	0.0068	0.0064	0.0066	0.0063	0.0063	0.0064	0.0062	0.0057	0.0067

Table 6: Performance of robust linear models on test set.

By considering only robust predictors, the models were able to improve their performance on the test set on almost every run. The best model out of these runs is the one with 11 lags achieving a MSE equal to 0.0057, a significant improvement with respect to the results obtained by the best previous models. In this best model, all predictors have a positive coefficient, indicating a positive correlation with the response variable, and 2 out of 4 are statistically significant, namely *CVX* and *COVID-19*. To visually examine the predictions made by this best model and its performance on the overall dataset, a plot containing both the predicted values and the ground truth can be inspected in Figure 7.

As it is possible to notice, the model is not quite able to capture the variance of the WTI spot price on both the train set and test set. On the train set, which was used to fit the model, the predictions follow the correct direction of most spikes of the ground truth, but on the overall they do not explain accurately the variation of the response variable, as also indicated by the low R^2 achieved equal to 0.041. On the test set, the model performs well

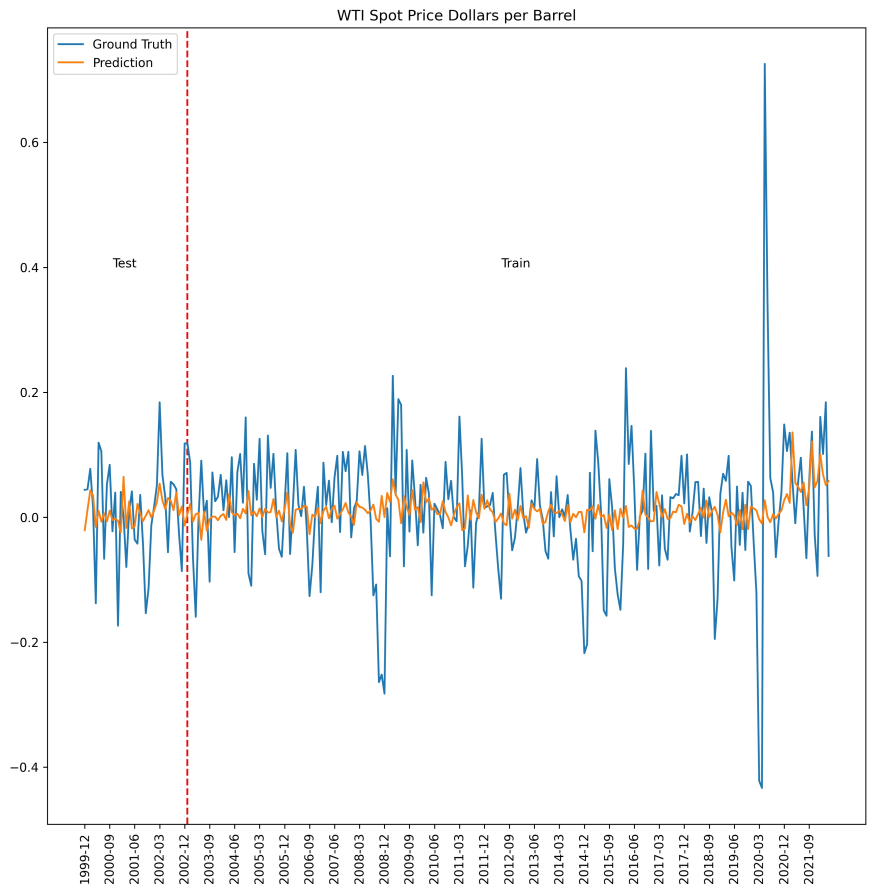

Figure 7: Visualization of the best econometric model performance on the entire dataset.

in terms of MSE and it predicts quite accurately the last section of data, but for the most part its predictions are not able to explain the ground truth trend.

5 Neural network analysis

Prediction of financial markets is often complicated due to non-linear relationships among variables. However, neural network architectures have recently shown promising results in their ability to model complex data dependencies and learn difficult patterns. For this reason, to investigate whether WTI spot prices can be forecast more accurately, another analysis was conducted using this time neural networks.

The data used during this analysis is the same used in the econometric analysis, thus considering also the conditional volatility of variables with ARCH effect extracted previously by the GARCH models. However, in order to take advantage of the properties of the LSTM, the data was processed in a different way. In particular, the data given in input to the model does not contain only one lag, but is processed as a sliding window of size equal to 12. In this way, the input is a sequence of 12 timesteps $(x_1, x_2, .., x_{t-1})$ and the neural network will be able to model the temporal dependencies in the data to forecast the WTI spot price (y_t).

The architecture of the neural network used in this analysis is illustrated in Figure 8.

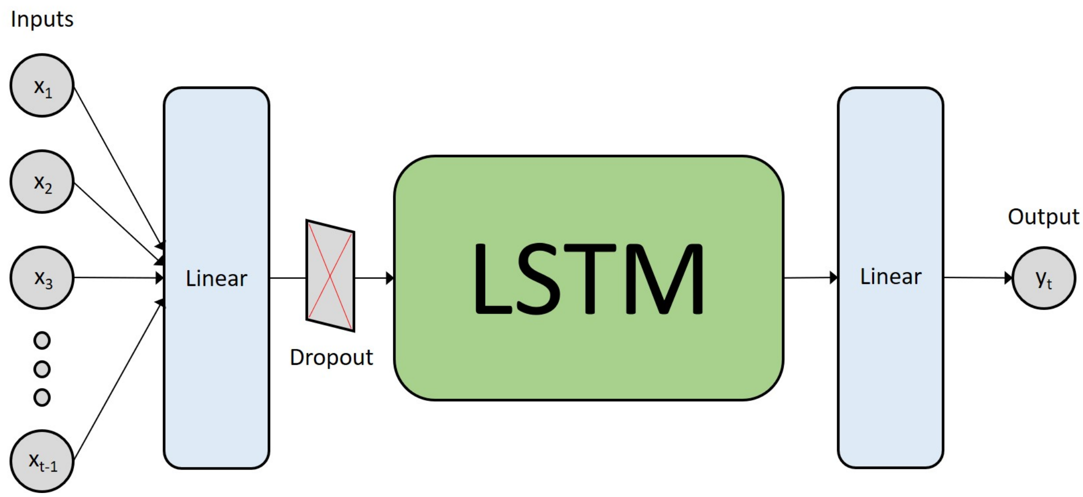

Figure 8: Neural network architecture.

The first component of the architecture is a Linear layer, also known as Fully-Connected layer, which is used to learn an initial representation of the data to fed into the LSTM. It takes in input the 12 timesteps composed of 35 variables and produces a sequence of the same length composed of their embeddings. Following, a Dropout layer [3] has been placed in order to introduce a regularization effect and prevent the neural network from overfitting the training data. Dropout layers randomly set the input units to 0 with probability p during training, while during inference no values are dropped, and have been shown to be effective in improving the generalization performance of neural networks. After this, the sequence is passed to an LSTM which is used to learn long-term dependencies and extract informative features out of the data. Lastly, the features computed by the LSTM, are passed to a final Linear layer which produces the output. In particular, the first Linear layer is composed of 16 neurons, the Dropout probability is set to 0.3, the LSTM has an hidden size of 16 and the final Linear layer is composed of 1 neuron.

The neural network was trained to optimize a MSE loss for 12000 epochs with a batch size of 32 samples, Adam optimizer [2] and a learning rate equal to 0.005. The training was performed on a GPU on Google Colab and a seed was set in order to make the results reproducibAt the end of each epoch of training, the model was also evaluated on the test set and its accuracy was recorded. The neural network achieved its best performance on the test at the 10265th epoch of training, with a MSE equal to 0.0042, thus smaller than the one achieved by the best econometric model. Moreover, in order to inspect the predictions made by the model and its performance on the overall dataset, a plot containing both the predicted values and the ground truth can be examined in Figure 9.

As it is possible to notice, the results achieved are significantly better than the ones obtained in the econometric analysis. On the train set, the neural network is able to perfectly fit the data, as expected given the small quantity of data which increases the probability of overfitting. However, it can be observed that the model also performs well on the test set, particularly due to the addition of the Dropout layer which promotes the generalization of the model. The predictions regarding WTI spot prices made by the neural networks accurately capture the variance of the response variable and its trend. Contrary to what occurred in

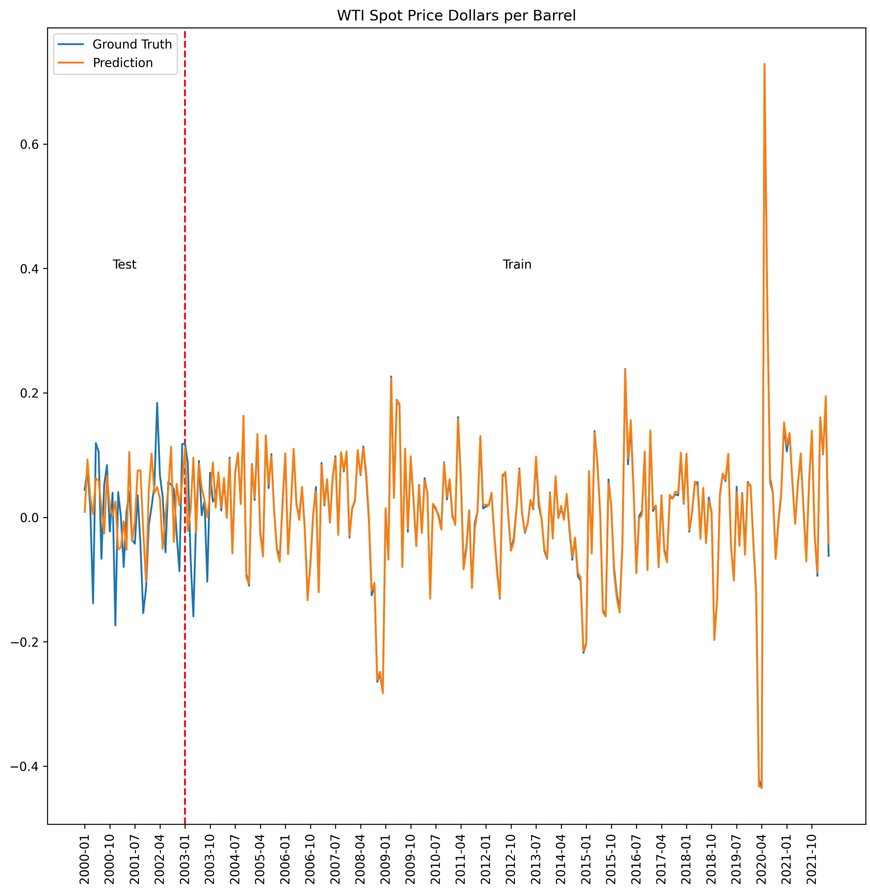

Figure 9: Visualization of the neural network performance on the entire dataset.

the econometric analysis, the low MSE achieved by the model is reflected in forecasts that accurately predict not only the direction but also the values of WTI spot prices.

6 Conclusions

The results obtained in this work showed that the neural network model outperformed the econometric models in forecasting WTI spot prices. The neural network was able to produce accurate predictions on the overall dataset and demonstrated a strong performance on the test set, achieving a lower MSE compared to the best econometric model.

Nonetheless, it is important to note that neural networks are quite difficult to train, particularly on small datasets, and do not provide any insights about why certain values are predicted. In contrast, econometric models can offer a more interpretable approach by explicitly modeling the relationships between different variables.

In conclusion, the use of neural networks can be a useful approach to forecast prices of financial assets. However, relying on these predictions carries a high level of risk, as prices are influenced by many factors, and neural networks, unlike econometric models, do not provide explanations for their predictions. As such, it may be advisable to consider both neural network and econometric models when making financial forecasts.

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[4] Wikipedia contributors. Energy policy and conservation act — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Energy_Policy_and_Conservation_Act&oldid=1125143848`, 2022.