 # Author: Christos Vasilopanagos

 # Date: 4 April 2025

Results and discussion

Models trained on the same datasets showed discrepancies in performance and overall fit as seen in their metrics, such as the mean squared error, mean absolute percentage error and the R-squared values. Specifically, the MLR model had the lowest metric values between all models. The model returned a MSE of 0.04, overall fit of 0.6 and a MAPE of 3.3%. Based on these values, the model qualifies within the acceptable range but bears limitations in terms of the variability observed for predicting crude oil inventory values. From a modelling perspective we attribute this to limited computational capacity, the linear relationship between dependent and independent variables as well as the observed multicollinearity among independent parameters (Haitovsky 1969). Upon testing for multicollinearity and the Variance Inflation Factor (VIF) between independent variables, the following was found (Alin, 2010) (Table 2):

Table 2. Variance inflation factors of features in MLR model

Variance Inflation Factors	
Feature	VIF
CL=F	10.8
USO	4.0

GSCI 10.1

Although just within the acceptable range, the model provides information on the relationship between the US crude oil inventory, oil and commodities markets in the selected region. The VIF value for WTI (CL=F) is approximately 10.83, indicating that there is high multicollinearity between WTI price and the other independent variables in the model. This suggests that the variance of the coefficient estimate for WTI is substantially inflated, indicating strong correlation between WTI and the other predictors (Narayan and Tagliarini, 2005). The United States Oil Fund (USO) VIF value was found at 3.98, which is below the threshold of 5. This indicates low to moderate multicollinearity between USO and the other independent variables. While there is some correlation between USO and the other predictors, it is not as strong as the correlation involving WTI. The Goldman Sachs Commodity Index (GSCI) had a VIF of 10.12 indicating a multicollinearity between GSCI index and the other independent variables. As with WTI, the variance of the coefficient estimate for GSCI index is substantially inflated, indicating strong correlation between GSCI index and the other predictors.

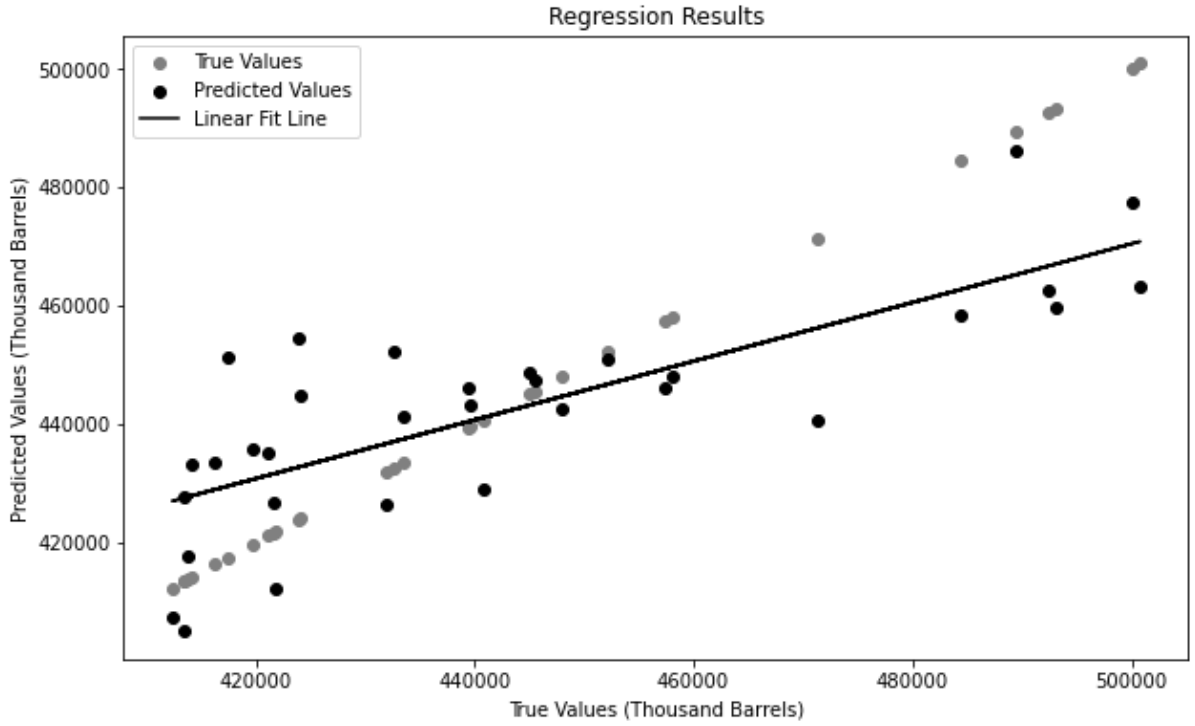

Fig 1. Multiple regression model of predicted and true values of US crude oil inventory in thousand barrels. Data were scaled to the range [0, 1] using Min-Max scaling.

The MLP performed a MAPE of 2.3, down by 30% from MLR, showing a significant improvement in performance. A MSE of 0.022 indicates a doubled goodness of the overall fit of the model, accompanied by a higher R-squared value. MLP's model performance indicates

the average squared difference between the predicted and actual crude oil inventory values is small and within acceptable range signifying a better model performance and suggesting that the MLP model has low prediction errors on average. Variance, i.e. the R-squared value of 0.72 indicates that approximately 72% of the variance in the crude oil inventory values can be explained by the input features included in the MLP model (GSCI, USO, and WTI historical market values). This suggests a moderate-to-strong level of predictive capability, indicating that the model captures a significant portion of the variability in the target variable. Fig 2 shows the training and validation loss curves of the MLP. The training loss shows a decreasing trend over epochs. This indicates that the model is learning from the training data and improving its performance without overfitting, particularly seen between 0 and 60 epochs. Consistent small gaps between the training and validation loss curves suggest that the model is not overfitting, i.e. the training loss is not lower than the validation loss. Validation loss decreases initially until the 60^{th} epoch, remains stable until the 80^{th} and then starts to increase slightly followed by a stabilization period. This indicates that the model is generalizing well to unseen data, indicating good generalization performance (Diaconsescu, 2008). The curves are smooth without significant fluctuations or irregularities. This indicates that the model is stable and not overly sensitive to small changes in the training or validation data. The curves reach a stable point after a certain number of epochs, namely between 80 and 85. This indicates that the model has converged and further training may not significantly improve performance.

The NARX model had a MAPE of 2.21% and a MSE of 0.02 indicating that, on average, the model's predictions deviate from the actual values by approximately 2.21% indicating that the model's predictions are close to the actual values. The R-squared score of 0.78 indicates that approximately 78% of the variance in the target variable (crude oil inventory values) can be explained by the input features used in the NARX model (Azizi, 2023). The plot (Fig 2) is demonstrating a strong fit for crude oil inventory values shows as high degree of concordance between the predicted and actual inventory levels over time. Overall, the results obtained from the NARX model suggest that it performs well and better than the other two in predicting crude oil inventory values (see Table 3)

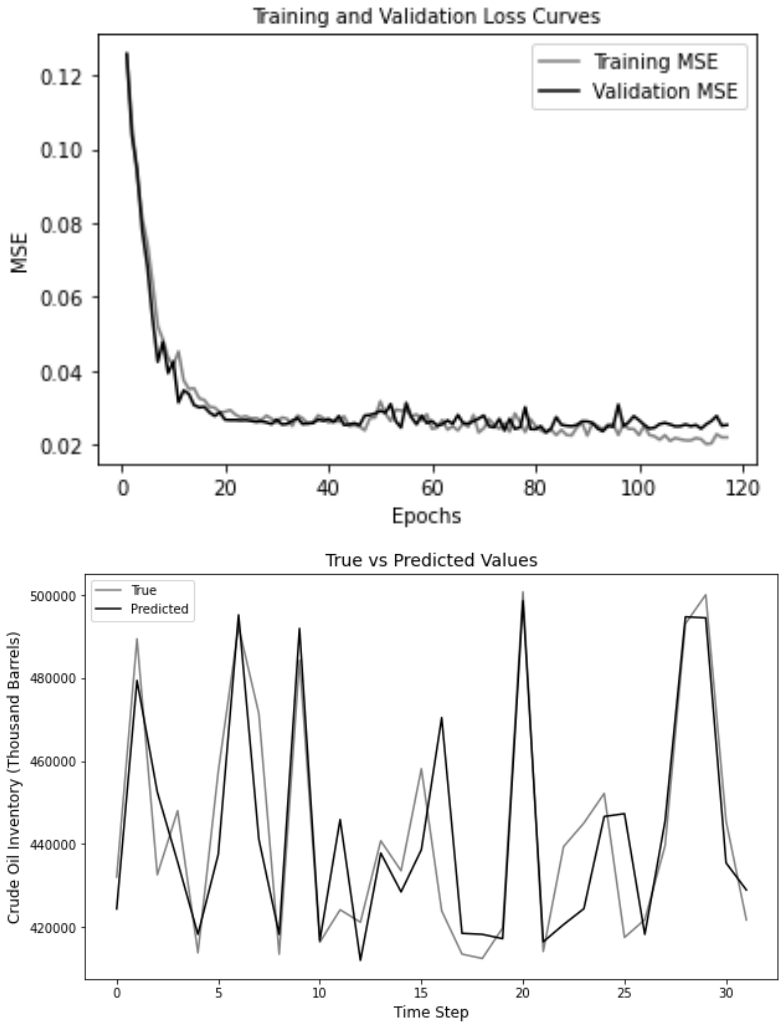

Fig 2. Up: Training and validation loss curves of the MLP model for a total of 117 epochs. Down: Comparison of True vs Predicted Crude Oil Inventory Values, using the NARX model. Data were normalized is the range $[0, 1]$. After normalization, the minimum value of each feature and the target variable was mapped to 0, and the maximum value to 1. All other values were linearly scaled between 0 and 1 based on their relative positions within the original range of the data.

Table 3. Model performance metrics

Models	MAPE (%)	MSE	R squared score
MLR	3.3	0.04	0.6
MLP	2.3	0.022	0.72
NARX	2.21	0.02	0.78

Table 4. Confidence Intervals for model coefficients for MLP and MLR Models.

	MLR confidence interval coefficients			MLP confidence intervals for layers	
	Low	High		Low	High
			Layer 1	-0.31	0.24
WTI (CL=F)	-1.32	0.03	Layer 2	-0.24	0.22
USO	-0.26	0.42	Layer 3	-0.32	0.30
GSCI	-1.06	0.24	Output layer	-0.53	0.57

A predict_with_uncertainty method with dropout during inference was employed to estimate prediction intervals and quantify uncertainty of the NARX model (Labach et al., 2019; . In total, 100 samples were used to assess the model's predictions with dropout applied during inference. This approach allowed for the estimation of both the mean and standard deviation of the predictions, providing insights into the model's expected performance and the level of uncertainty associated with each prediction. By capturing the variability inherent in the model's predictions, the method enabled a probabilistic assessment of forecast reliability, which is crucial for decision-making under uncertainty. The Mean of STD applied in the NARX model was calculated to be 8.1549×10^{-8}. The mean standard deviation (MSD) serves as a measure of the average uncertainty observed across all predictions generated by the NARX model. A low MSD suggests minimal variability or uncertainty in the model's predictions, indicating consistent outcomes for diverse input scenarios. However, while reduced variability is often favorable, particularly in scenarios requiring precise and reliable predictions, it does not uniformly guarantee the model's effectiveness.

The comparison of the predictive models, including Multiple Linear Regression (MLR), Multilayer Perceptron (MLP), and Nonlinear Autoregressive with Exogenous Inputs (NARX), reveals distinct performance dynamics. While MLR represents a baseline model, MLP and

NARX introduce increasing complexity. Across multiple evaluation metrics such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and R squared score, NARX consistently outperforms MLP and MLR. This superiority is evidenced by lower MAPE and MSE values and a higher R squared score, indicative of better predictive accuracy and explanatory power. The observed improvement trend from MLR to MLP to NARX underscores the significance of model sophistication in enhancing predictive capabilities. Additionally, NARX demonstrates superior generalization ability, suggesting its potential for robust performance on unseen data. Consequently, while MLR may suffice for basic modelling tasks, NARX emerges as a preferred choice for predictive modelling, particularly in scenarios demanding high accuracy and generalization as is the case with crude oil trading.

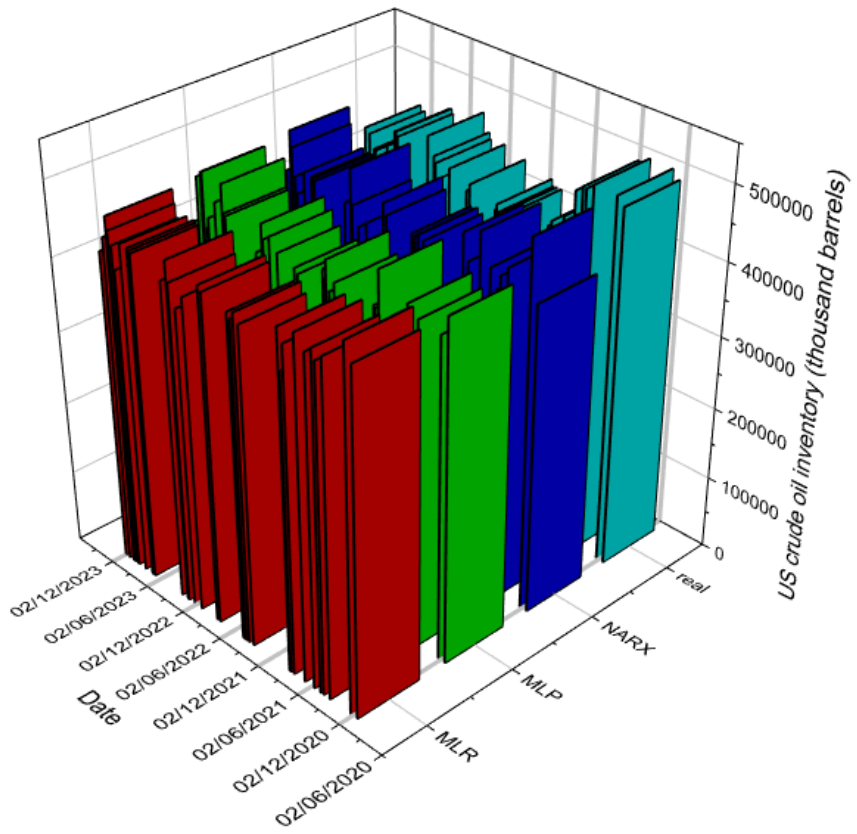

Fig 4. Real and predicted values of US crude oil inventory as predicted by MLR, MLP and NARX models respectively.