geospatial

Integration of Multi-Frequency Ensemble Machine Learning Models

Towards Crude Oil Price Forecasting

Farid Javadnejad

Dec 2022

Index

- Introduction
- Methodology
  - Datasets
  - Data Wrangling
  - Exploratory Data Analysis
  - Training
  - Modelling
- Results and Discussion
- Future Work

Image: wallpaperflare.com

Introduction

Context

- Can we predict the crude oil price prediction for the next 6 months using historical data of oil prices and other socio-economic features with 99% accuracy?
- Crude oil and other refined liquid products from fossil fuels are critical contributors to the world economy.
- Crude oil prices are difficult to predict accurately due to the number of influencing factors and their complex behaviour
- Accurate oil price forecasting is crucial for decision-making support for the manufacturing, logistics, and government sectors to guide industrial and social policies and practices

Source: www.istockphoto.com

Introduction

Criteria for success

- The project's scope is to use data science techniques to achieve 99% accuracy in predicting short-term oil price trends.

Solution space

- Predicts oil price for the next 6 month

Solution Constraints

- Use the West Texas Intermediate (WTI) as benchmark
- Limit the data from 2000 to 2022.

Key data sources

- EIA
- FRED
- WSJ
- Investing

Methodology

Datasets

- Factors influencing the crude oil market dynamic:
  - Supply and demand
  - Financial markets
  - Policies & Politics
  - Global events
  - Alternative resources
  - Development technologies
  - Consumption patterns

Category	Variable	Unit	Source
Crude Oil Price	West Texas Intermediate Crude Oil Price	dollars per barrel	EIA
Supply	Crude Oil Production, Total OPEC	million barrels per day	EIA
	Crude Oil Production, Total non-OPEC	million barrels per day	EIA
	Crude Oil Production, NGPL, and other liquids production, World	thousand barrels per day	EIA
	Crude Oil Production, U.S.	million barrels per day	EIA
Replacement Cost	Henry Hub Natural Gas Spot Price	dollars per million btu	EIA
Demand	Liquid Fuels Consumption, Total OECD	million barrels per day	EIA
	Liquid Fuels Consumption, Total non-OECD,	million barrels per day	EIA
	Federal Funds Effective Rate	percent, not seasonally adjusted	FRED
	Index of Global Real Economic Activity	index, not seasonally adjusted	FRED
	US Consumer Price Index (CPI): All Commodities	index, 1982-1984=1.00	EIA
	US Consumer Price Index (CPI): Energy for the United States	index 2015=100, not seasonally adjusted	FRED
	US Producer Price Index (PPI): All Commodities	index, 1982=1.00	EIA
	US Producer Price Index (PPI): Petroleum	index, 1982=1.00	EIA
	Producer Price Index (PPI) of Euro Area (19 Countries)	index 2015=100, not seasonally adjusted	FRED
	US Manufacturing Production Index (PMI)	index, 2017=100 (seasonally adjusted)	EIA
Inventory	Petroleum Inventory, Total OECD	million barrels, end-of-period	EIA
	Petroleum Inventory, U.S. Total	million barrels, end-of-period	EIA
	US Crude Oil Inventory: Strategic Petroleum Reserve (SPR)	million barrels, end-of-period	EIA
	US Crude Oil Inventory: Non-SPR	million barrels, end-of-period	EIA
Monetary Market	Real Broad Dollar Index	index Jan 2006=100, not seasonally adjusted	FRED
	US Dollar Index (DXY)	index	Investing
	U.S. Dollars to Euro Spot Exchange Rate (DEXUSEU)	U.S. dollars to one euro, not seasonally adjusted	FRED
Stock Market	S&P 500 Index	index	WSJ
	Dow Jones Industrial Index	index	WSJ
	NASDAQ index	index	WSJ
Commodity Market	Gold Futures Historical Data	dollar per ounce	Investing
	Copper Futures Historical Data	dollar per pound	Investing
Policy Uncertainty	Global Economic Policy Uncertainty Index: Current Price Adjusted GDP	index, not seasonally adjusted	FRED
Technology	Refiner Wholesale Gasoline Price	cents per gallon	EIA
	Diesel Fuel Refiner Wholesale Price	cents per gallon	EIA
	Brent Crude Oil Spot Price	dollars per barrel	EIA

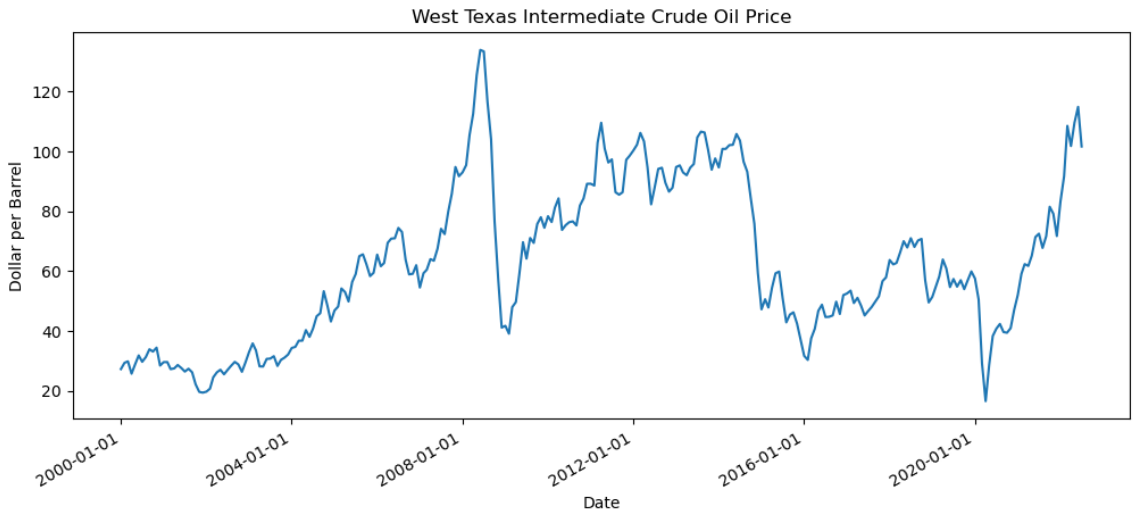
Methodology

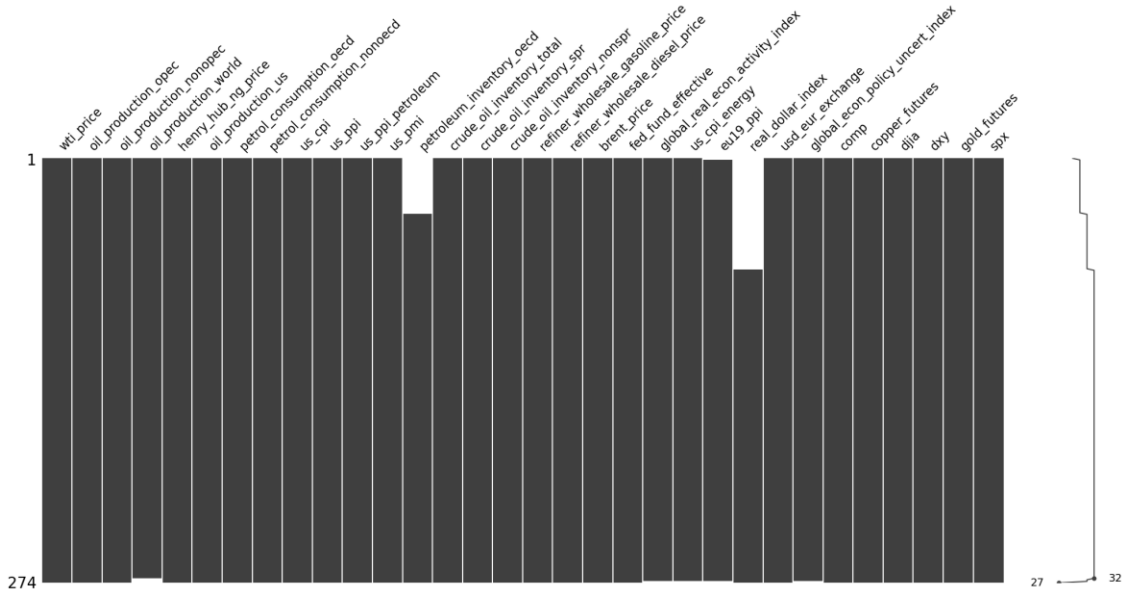
Data Wrangling

Missing Values

- Used Pandas is for Python programming language.
- The final dataset sized (274, 32) that represent 274 months for 32 feature variables.
- After treating missing data, the final dataframe sized (271, 30).

Variable	Count	Percentage
oil_production_world	3	1.1%
petroleum_inventory_oecd	36	13.1%
global_real_econ_activity_index	1	0.4%
us_cpi_energy	1	0.4%
eu19_ppi	2	0.7%
real_dollar_index	72	26.3%
global_econ_policy_uncert_index	1	0.4%





Methodology

Exploratory Data Analysis

Feature Engineering

- WTI-Brent Spread is difference between Brent and WTI prices.
  - Difference between the prices often reflect technical, supply/demand or geopolitical issues.

- The crack spread is the price difference between crude oil and its refined oil
  - Reflectis the supply and demand relationship between the crude oil market and its refined product market

- WTI-Brent spread = WTI spot price - Brent spot price

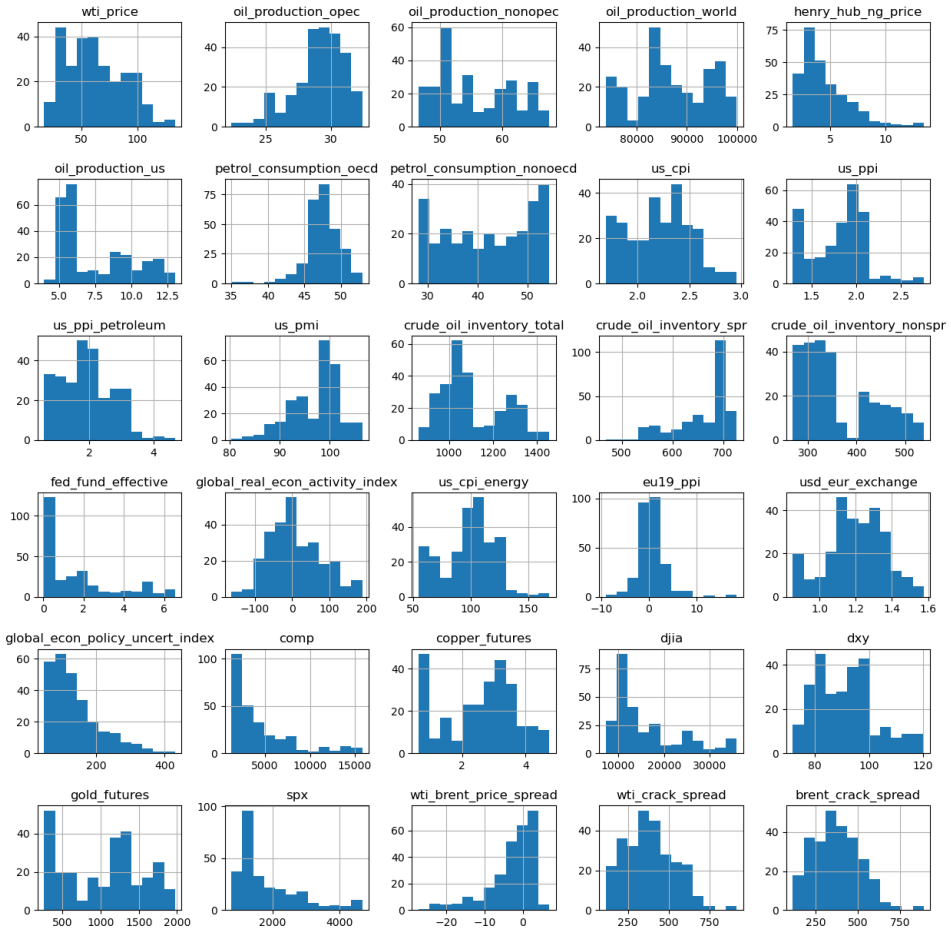
- WTI crack spread = 3 × WTI spot price – 2 × Gasoline Price – 1 × Diesel Fuel Price

- Brent crack spread = 3 × Brent spot price – 2 × Gasoline Price – 1 × Diesel Fuel Price

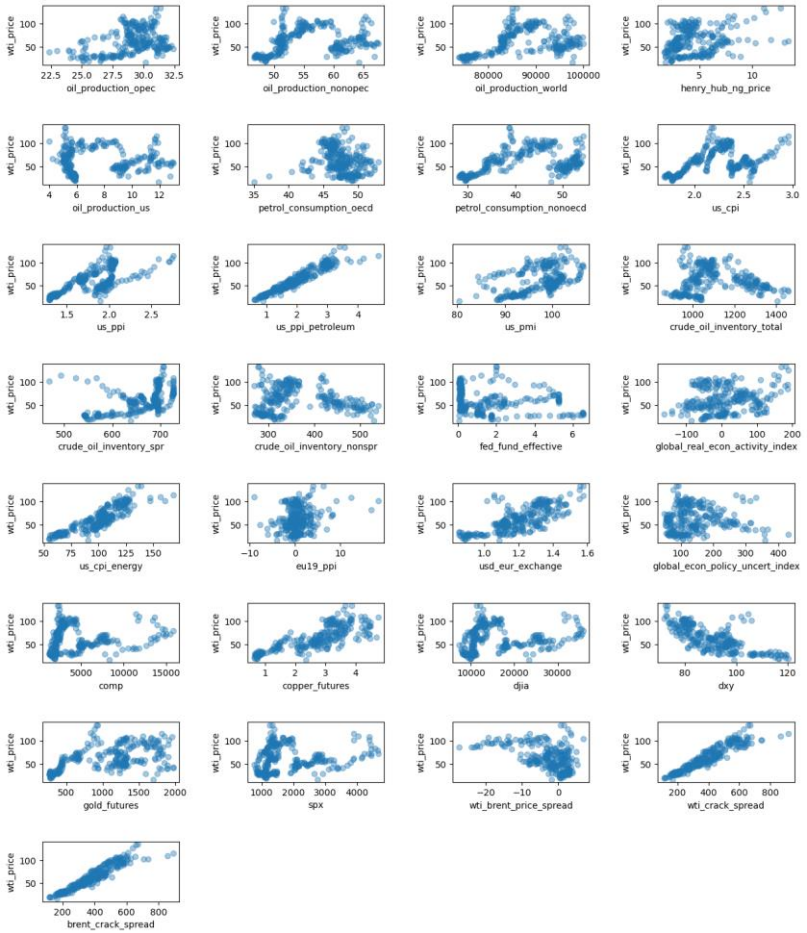
Methodology

Exploratory Data Analysis

Histograms



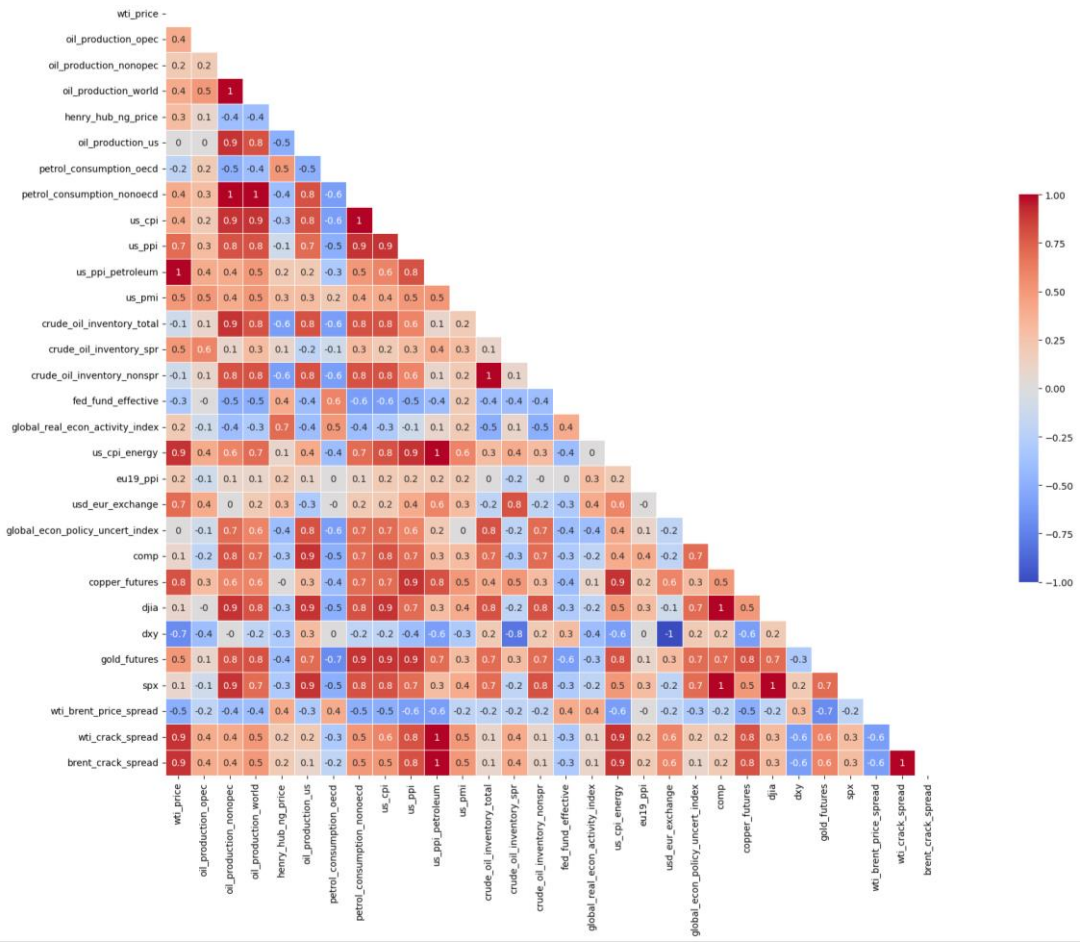
Scatter plots



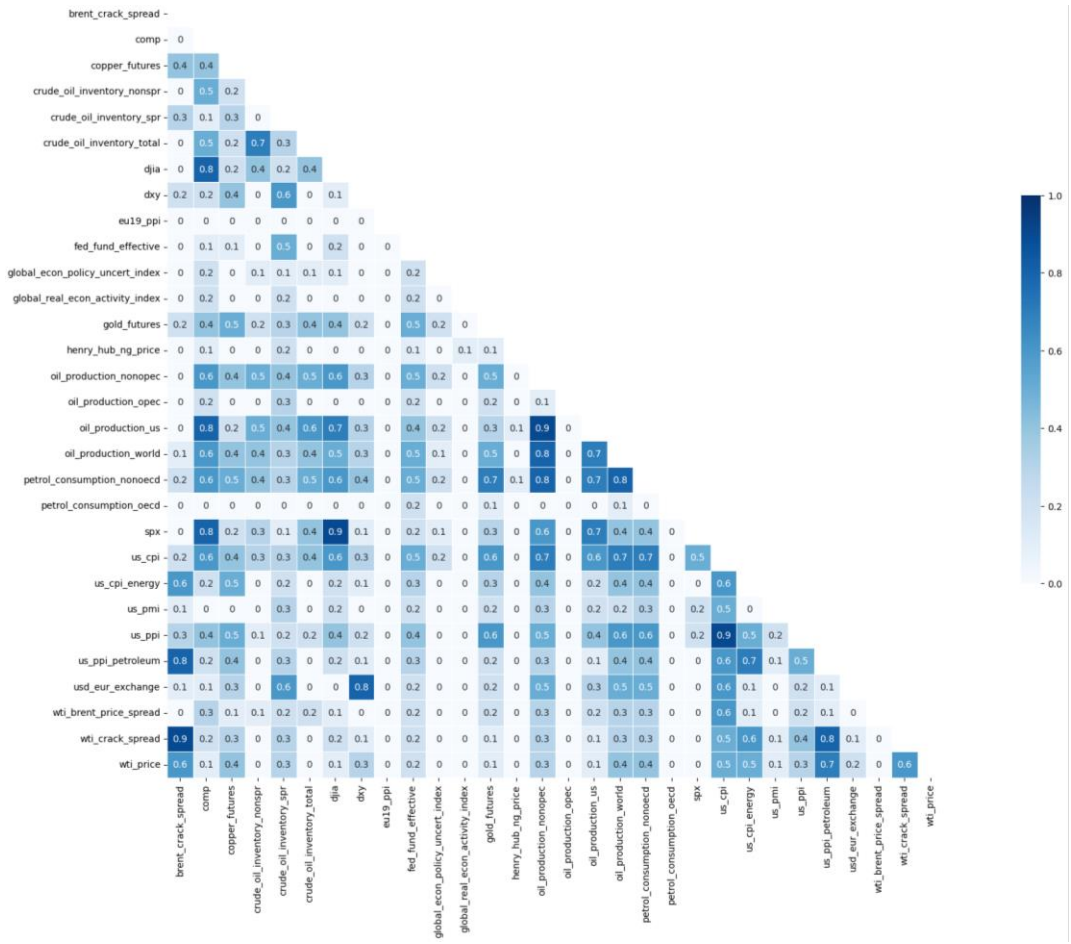
Methodology

Exploratory Data Analysis

Correlation coefficients matrix



Predictive Power Score (PPS) matrix



Methodology

Pre-processing and Training

Time shift

- The objective is to predict the dependent variable using the independent variables in the future, where the independent variables are also unknow.
- A data preparation step is to shift the dependent values in time to be associated with the independent variables of previous time period

$$Y_t = f(X_{t-n}, \beta_t) + e_{t-n}$$

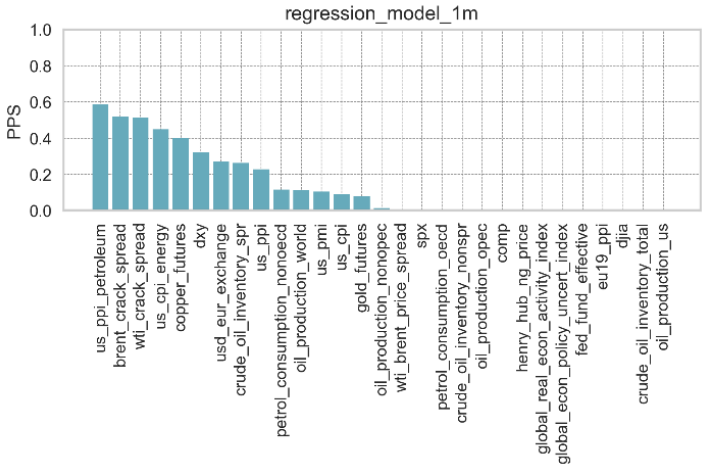
- We performed time shifting for frequency of $n = \{1, 2, 3, 4, 5, 6\}$ months

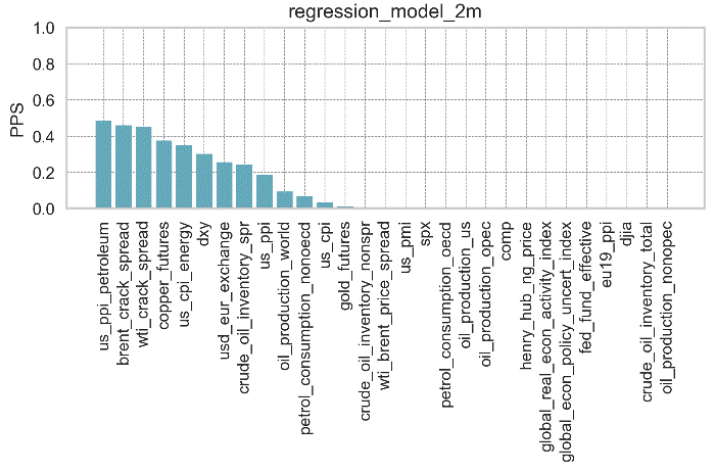
- After performing time shifting, the dependent variable will have a new set of independent variables that are specific to that frequency.
- In each frequency a different variable may have stronger correlation in the WTI Price.
- In real world scenario, this can be described and the time-lag for an effect to make an impact on the WTI prices.

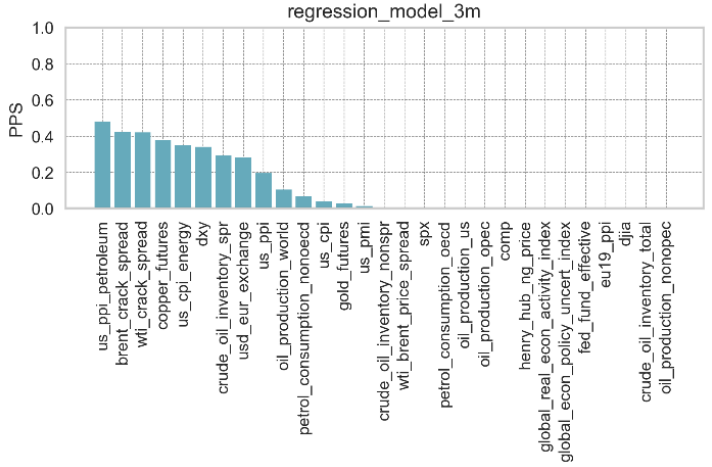
Methodology

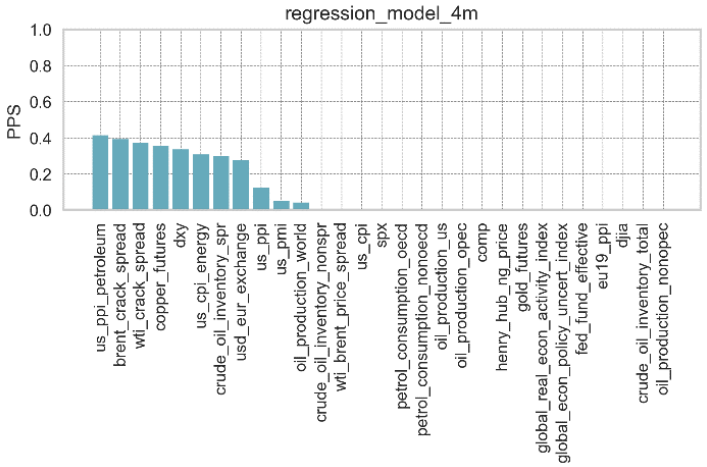
Pre-processing and Training

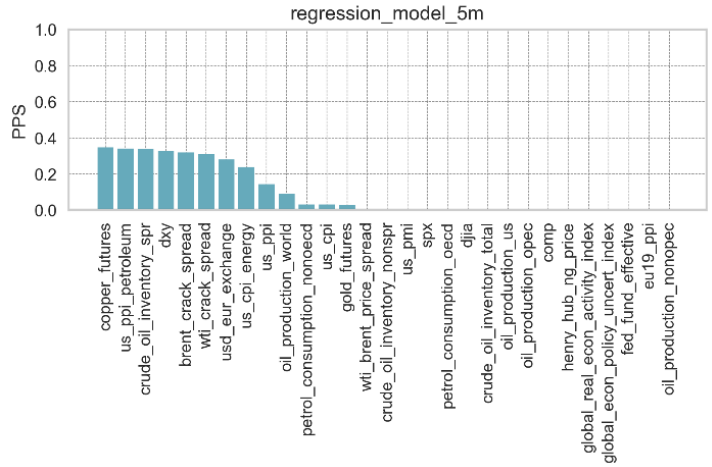
Multi-frequency scenarios

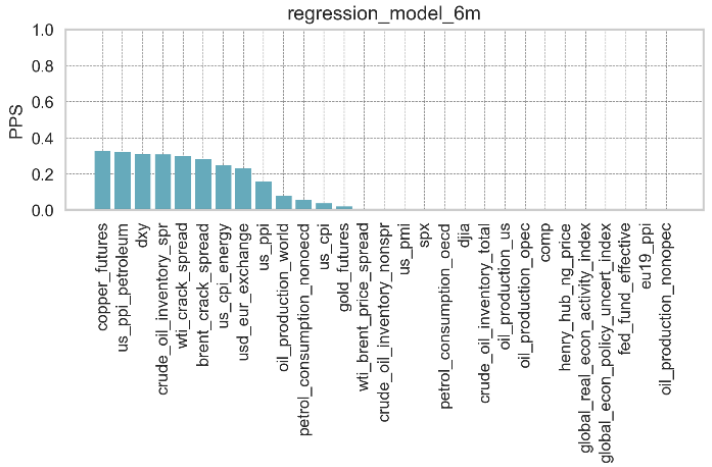












Methodology

Pre-processing and Training

Training

- Use PyCaret models to construct and deploy the models.
  - an open-source low code Python library that automates machine learning (AutoML)
- The library manages twenty-five different algorithms for regression, such as
  - Extra Trees Regressor (E.T.)
  - Gradient Boosting Regressor (GBR)
  - Extreme Gradient Boosting Regressor (XGB)
  - Random Forest Regressor (R.F.)
  - Linear Regression (L.R.),
  - AdaBoost Regressor (ADA),
- It also incudes eighteen other algorithms for classification.

PyCaret regression session

Description	Value
Session id	786
Target	wti_price
Target type	Regression
Data shape	(270, 30)
Train data shape	(188, 30)
Test data shape	(82, 30)
Numeric features	29
Preprocess	True
Imputation type	simple
Numeric imputation	mean
Categorical imputation	constant
Low variance threshold	0
Transformation	True
Transformation method	yeo-johnson
Normalize	True
Normalize method	zscore
Fold Generator	KFold
Fold Number	10

Methodology

Pre-processing and Training

Training

Performance of ML models for 1-month frequency scenario

	Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
et	Extra Trees Regressor	4.51	40.47	6.32	0.95	0.12	0.08
gbr	Gradient Boosting Regressor	5.10	48.24	6.85	0.94	0.12	0.09
lightgbm	Light Gradient Boosting Machine	5.22	53.37	7.25	0.93	0.12	0.09
rf	Random Forest Regressor	5.30	55.06	7.38	0.93	0.13	0.10
lr	Linear Regression	5.90	56.68	7.49	0.93	0.16	0.12
ada	AdaBoost Regressor	5.80	59.57	7.63	0.92	0.13	0.11
huber	Huber Regressor	5.81	58.89	7.64	0.92	0.16	0.12
ridge	Ridge Regression	5.94	59.68	7.69	0.92	0.16	0.12
br	Bayesian Ridge	6.02	62.60	7.88	0.92	0.16	0.12
lasso	Lasso Regression	6.26	68.05	8.22	0.91	0.16	0.12
knn	K Neighbors Regressor	5.89	73.74	8.42	0.90	0.14	0.11
omp	Orthogonal Matching Pursuit	6.31	75.90	8.70	0.90	0.17	0.12
en	Elastic Net	6.86	79.07	8.86	0.89	0.17	0.13
dt	Decision Tree Regressor	6.64	84.37	9.07	0.88	0.15	0.12
par	Passive Aggressive Regressor	8.66	132.63	10.90	0.82	0.31	0.19
llar	Lasso Least Angle Regression	12.88	249.70	15.76	0.67	0.28	0.26
dummy	Dummy Regressor	23.21	756.93	27.45	-0.01	0.48	0.50

Methodology

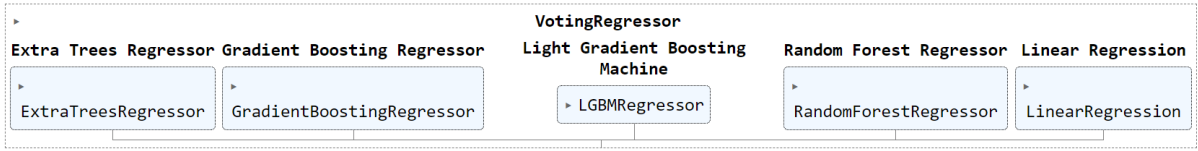
Pre-processing and Training

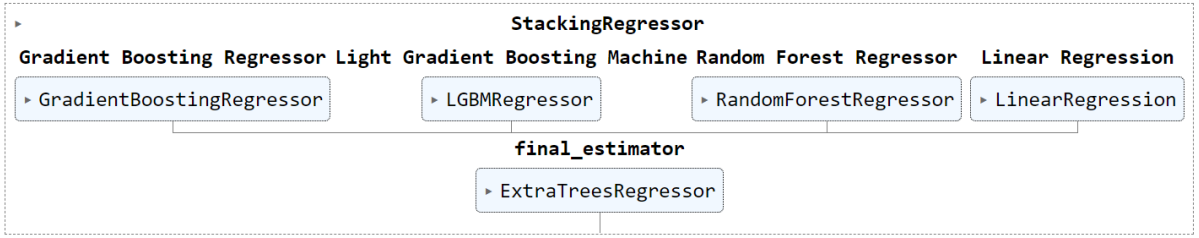
Hyper-parameterization

- Dynamically optimized the top 5 models based on RMSE value through 120 iterations with 5-fold CV (totaling 600 fits).
- The optimized top five models for each time-frequency and build ensemble models
- Ensemble methods improve the accuracy of the integrated model
- The most common ensemble methods are voting regressor and stacking regressor.

Ensemble Models

- Voting regressor uses a majority vote to build consensus of final prediction values.
- Stacking uses meta-learning to create multiple base estimators to generate the final prediction

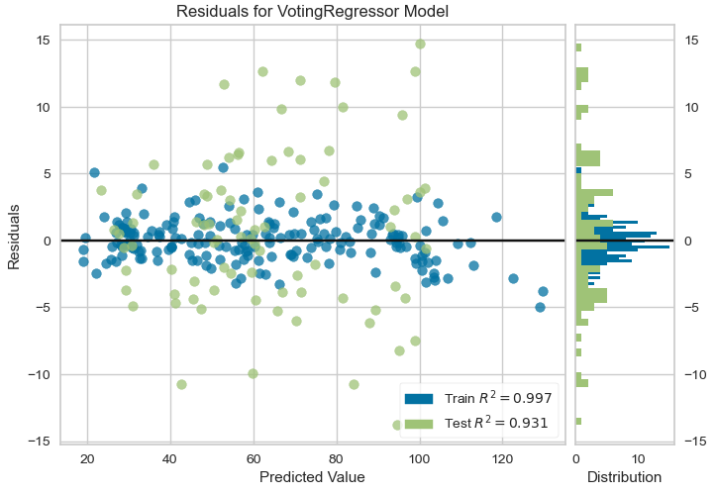


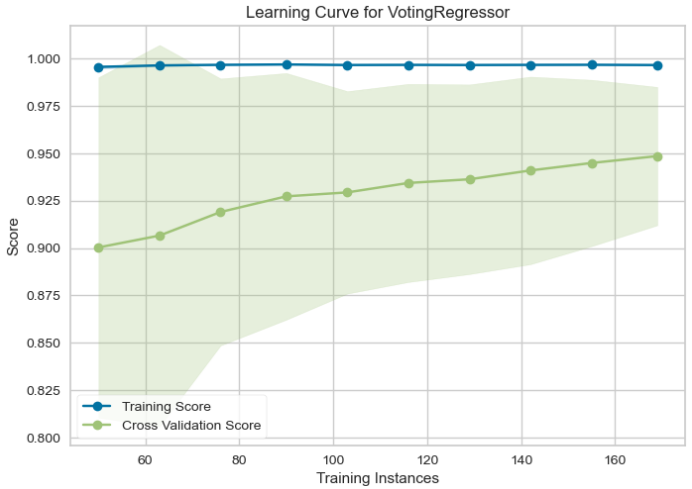


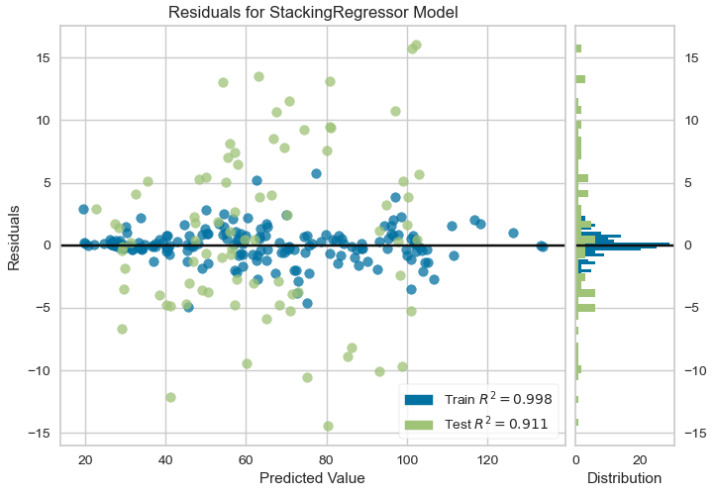
Methodology

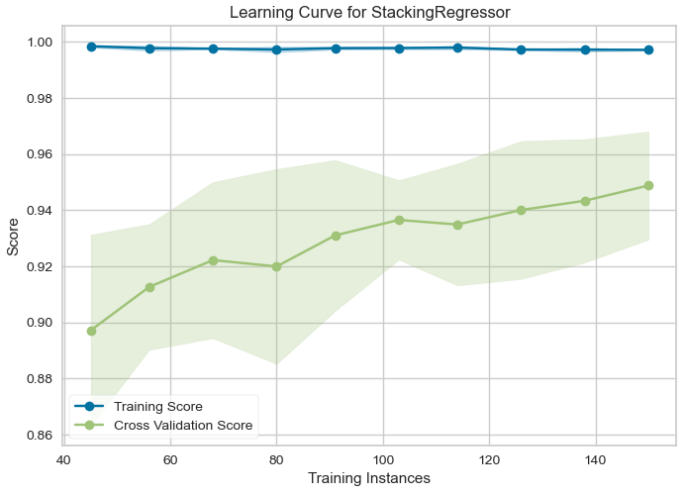
Pre-processing and Training

Ensemble model performance









Ensemble model	Metric	MAE	MSE	RMSE	R2	RMSLE	MAPE
Voting Regressor	Mean	4.5984	38.92	6.187	0.948	0.112	0.086
	Std	0.538	10.08	0.802	0.01	0.034	0.023
Stacking Regressor	Mean	4.5974	39.78	6.157	0.947	0.110	0.084
	Std	0.813	17.59	1.369	0.021	0.041	0.028

Methodology

Modelling

Final ensemble models

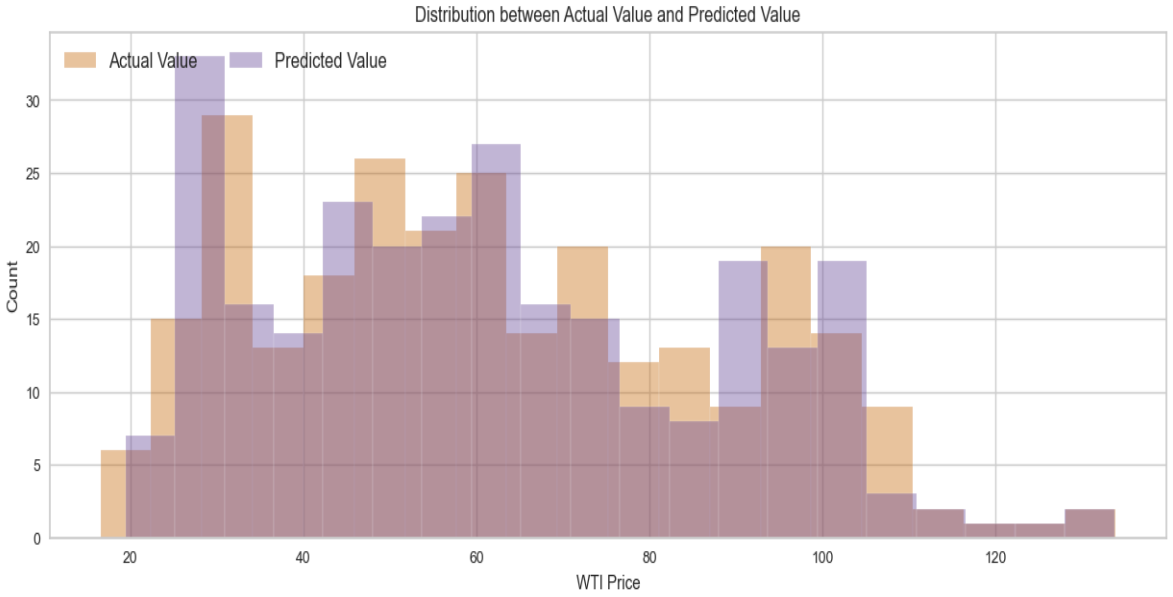
Model	Ensemble	MAE	MSE	RMSE	R2	RMSLE	MAPE
1-month	Stacking Regressor	0.9828	1.9863	1.4094	0.9971	0.0246	0.0164
2-month	Voting Regressor	1.4974	4.0136	2.0034	0.9941	0.0385	0.0273
3-month	Stacking Regressor	1.3449	4.1664	2.0412	0.9938	0.0344	0.0224
4-month	Voting Regressor	1.8882	6.3462	2.5192	0.9905	0.0503	0.0346
5-month	Stacking Regressor	1.2967	5.0315	2.2431	0.9925	0.0553	0.025
6-month	Stacking Regressor	1.2893	5.3536	2.3138	0.9920	0.0585	0.0253

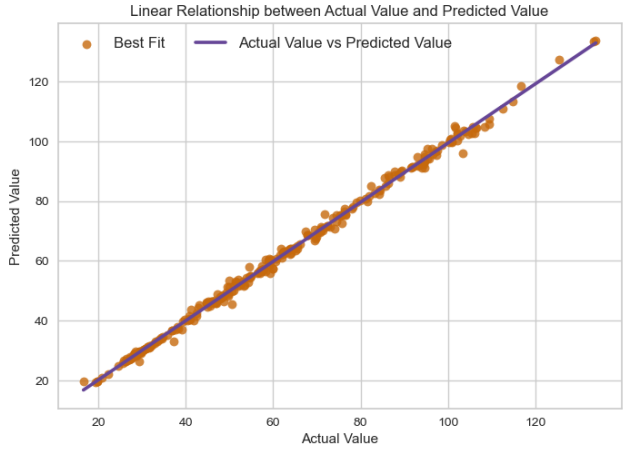
Methodology

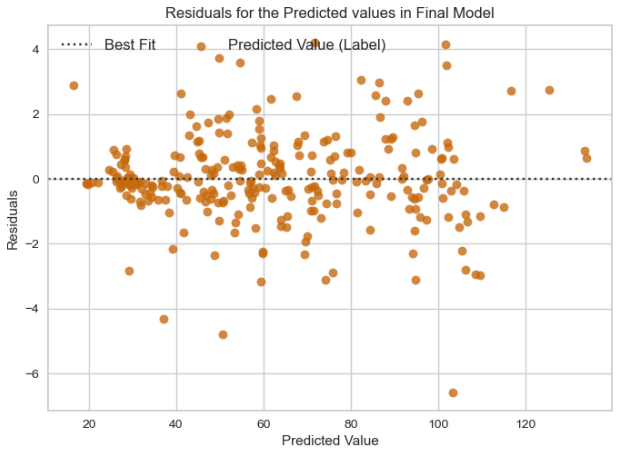
Modelling

Final ensemble models

The performance of final model for 1-month scenario



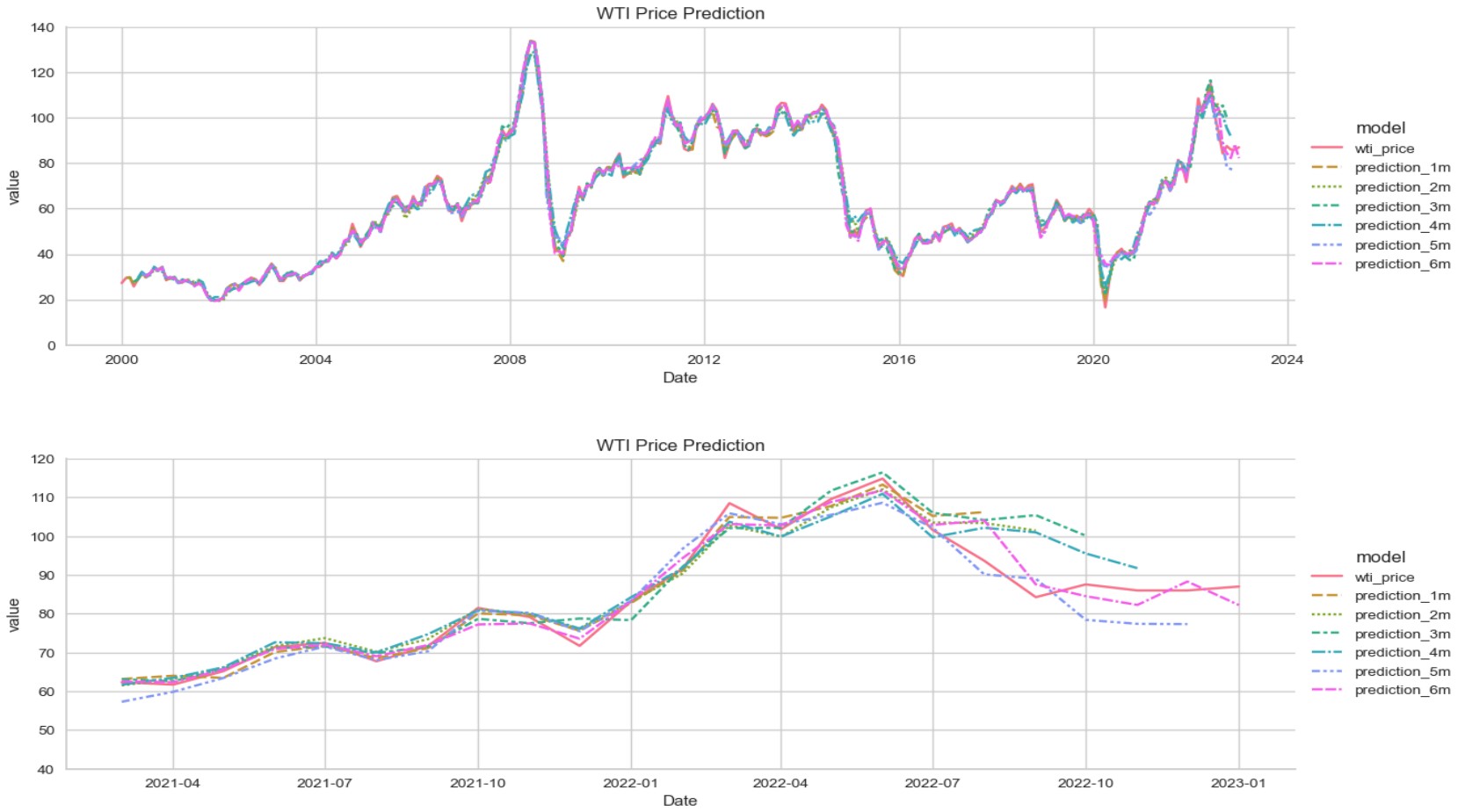




Methodology

Modelling

Final ensemble models



Methodology

Modelling

Integration into final model

- Fluctuations in multiple factors may influence the crude oil market at different frequencies.
- To account for this multi-frequency impact, we developed multi-frequency modeling approach.
- Here, we use the multiple predicted values for future prices for the crude oil and calculate a weighted average prediction value

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

$$w = 1/\overline{RMSE}$$

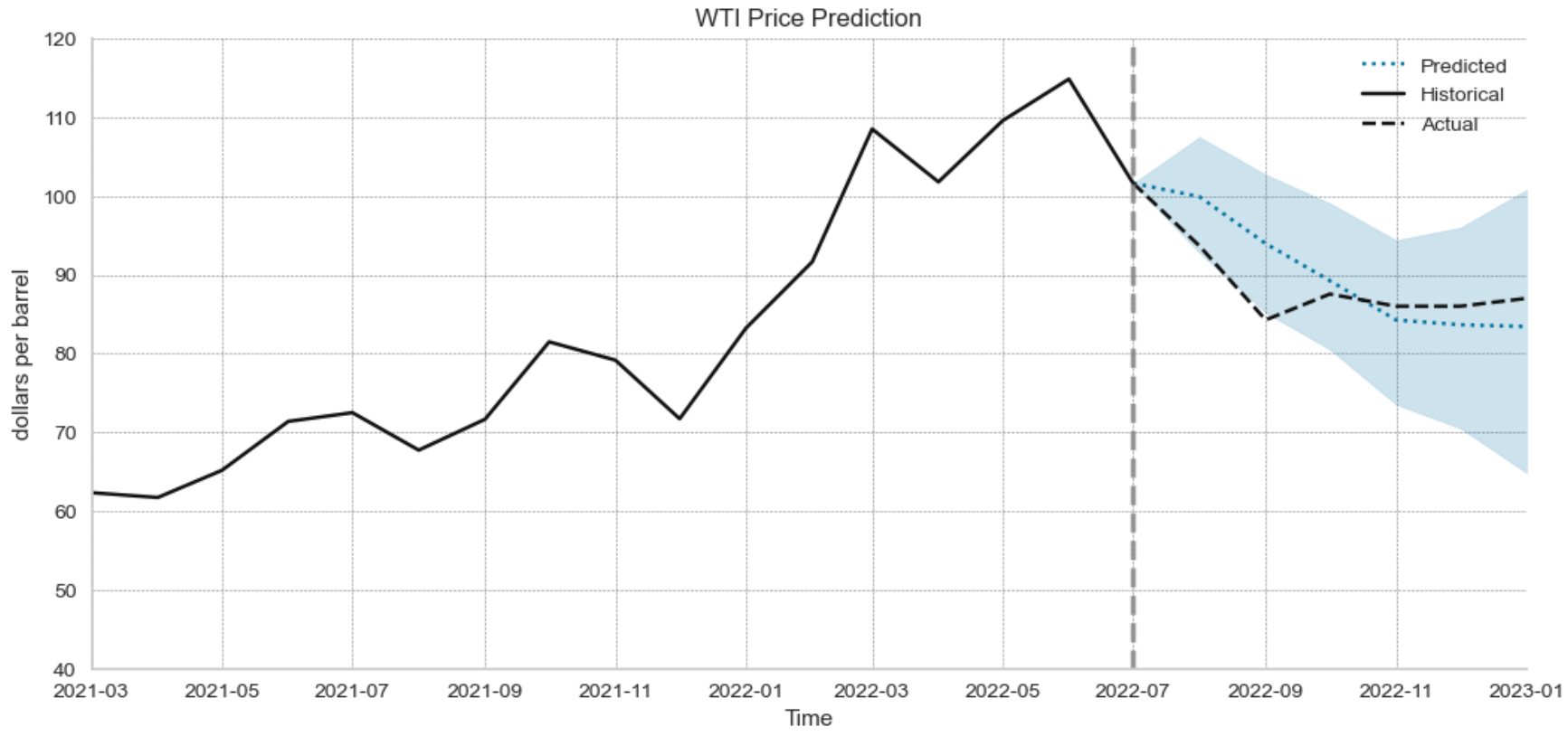
$$\overline{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} RMSE_i{}^2}{n}}$$

$$\text{C. I.} = \pm z_{1-0.01/2} \times RMSE$$

Methodology

Modelling

Final model & predictions



Results and Discussion

- An end-to-end workflow for predicting the WTI crude oil pricing using AutoML
- Data from publicly available resources
  - 33 features
  - Timeline Jan 2000 - December 2022.
  - After cleaning 32 feature for 274 months.
- Exploratory data analysis studied associations between features.
- Three new features were created via feature engineering.
- PyCaret was used to pre-process, split test/train sets, and normalize the data.
- PyCaret was used to build, deploy and evaluate AutoML models

- 1- to 6-month multi-frequency models WTI price prediction was generated.
- For each scenario the performance was evaluated by hyper-parametrization, and the top 5 performance models were stored
- The top 5 models were used to build ensemble models.
- The weighted average values of multiple-frequency predicted WTI prices were calculated and plotted with lower and upper bound conference internals.
- Plots the actual WTI against the predicted values shows that the model is successful in forecasting the oil price.

Results and Discussion

- For future studies, it is recommended to use other machine learning method such as
  - Long short-term memory (LSTM)
  - Random walk (RW)
  - Autoregressive integrated moving average models (ARMA),
  - Elman neural Networks (ENN),
  - ELM Neural Networks (EL),
  - Generalized regression neural network (GRNN)
- Then compare to the presented approach.