GOLD RETURN PREDICTION USING MACHINE LEARNING METHODS

GROUP 8

BIGDATA IN FINANCE II

INTRODUCTION

Existing literature

- Numerous studies on gold price prediction but few on return
- Technical features and traditional commodities features such as energy and material
- Limited attention to variable importance, especially in non-parametric models

The project

- Focus on gold return prediction
- Includes rarely explored features such as agricultural commodities
- Focus on economical variable importance

Project pipeline

Exploratory Data Analysis

Model Definition

Predictive Modelling

Performance Evaluation

Variable Importance

DATA

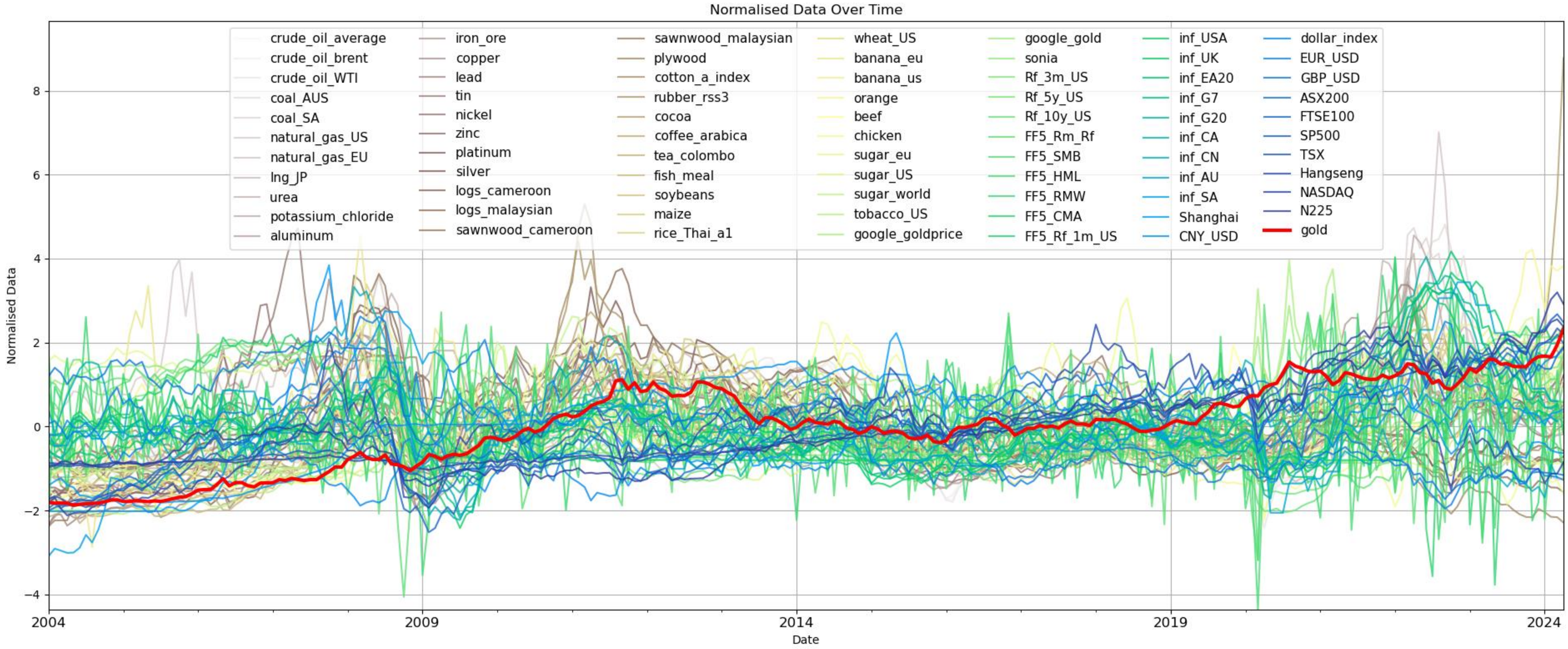
Metadata

- Time: 01-01-2004 → 01-02-2024, 242 monthly observation
- Number of features: 77
- Sources: OECD, The World Bank, Federal Reserve Bank of St. Louis, etc.

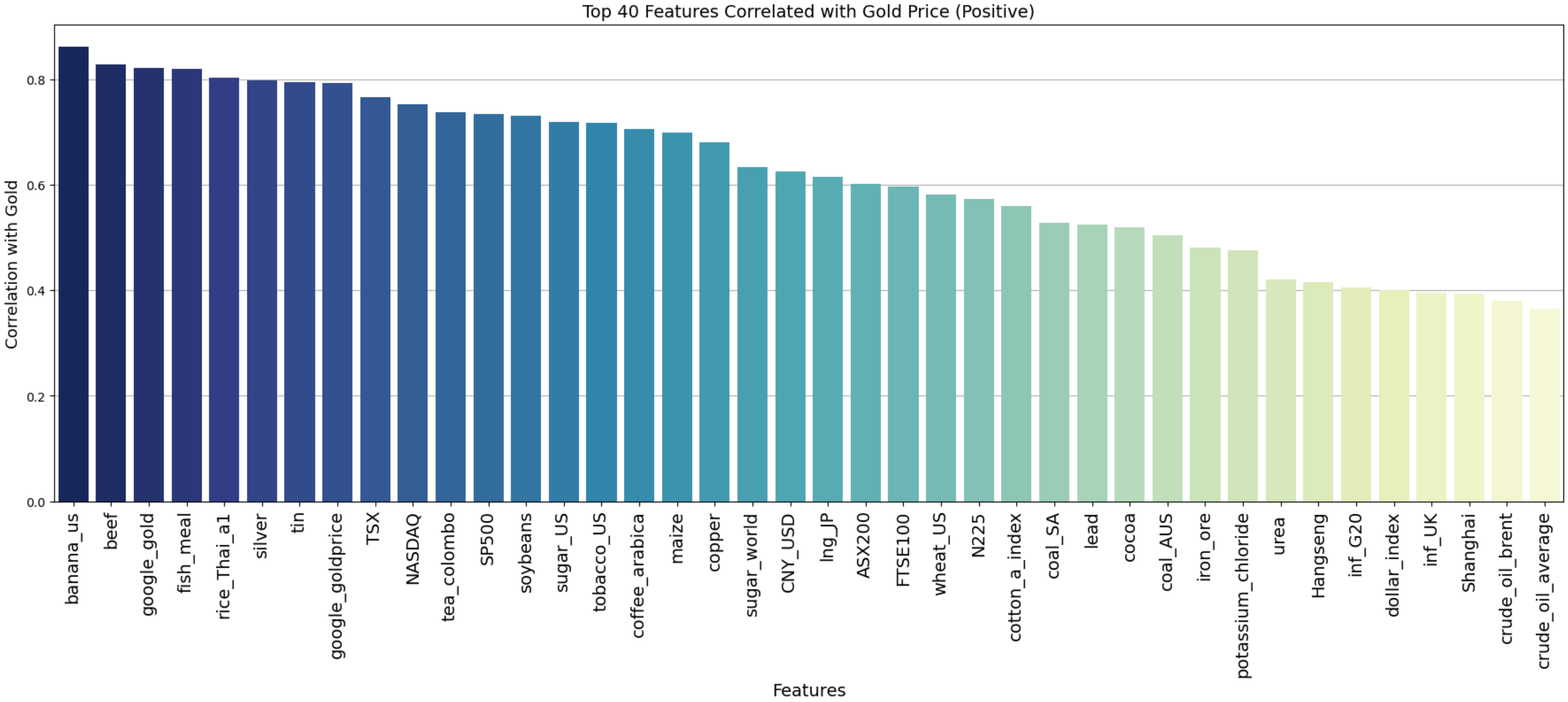
Features

Commodity Features	Market Features	Macroeconomic Features
Material Commodities	Exchange Rates	Dollar index
Agricultural Commodities	Index Funds	G7 Inflation
Resource Commodities	Fama French 5	G20 Inflation
	Interest Rate	Google Search Trends
	…	…

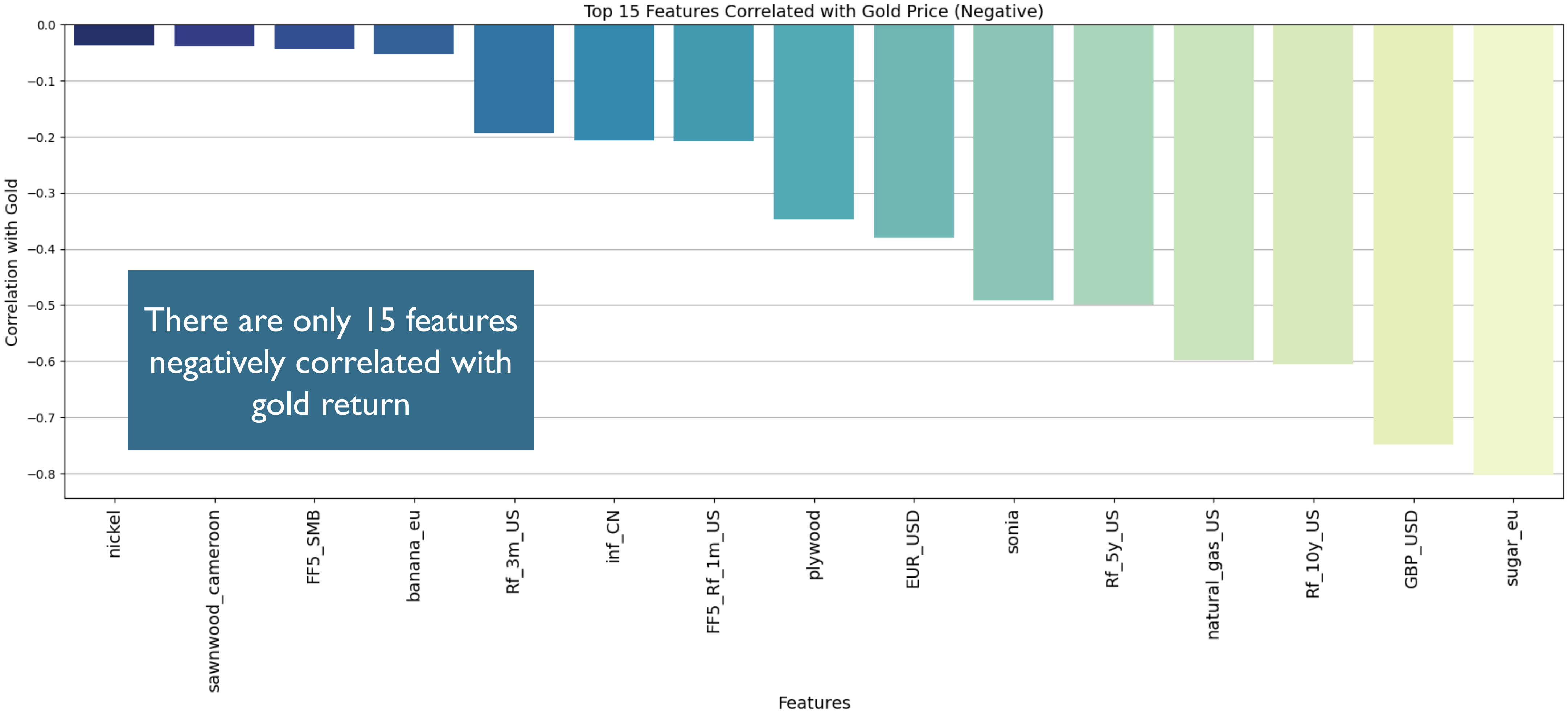
NORMALIZED PRICE DATA OVER TIME



PRICE CORRELATION (TOP 40 POSITIVE)

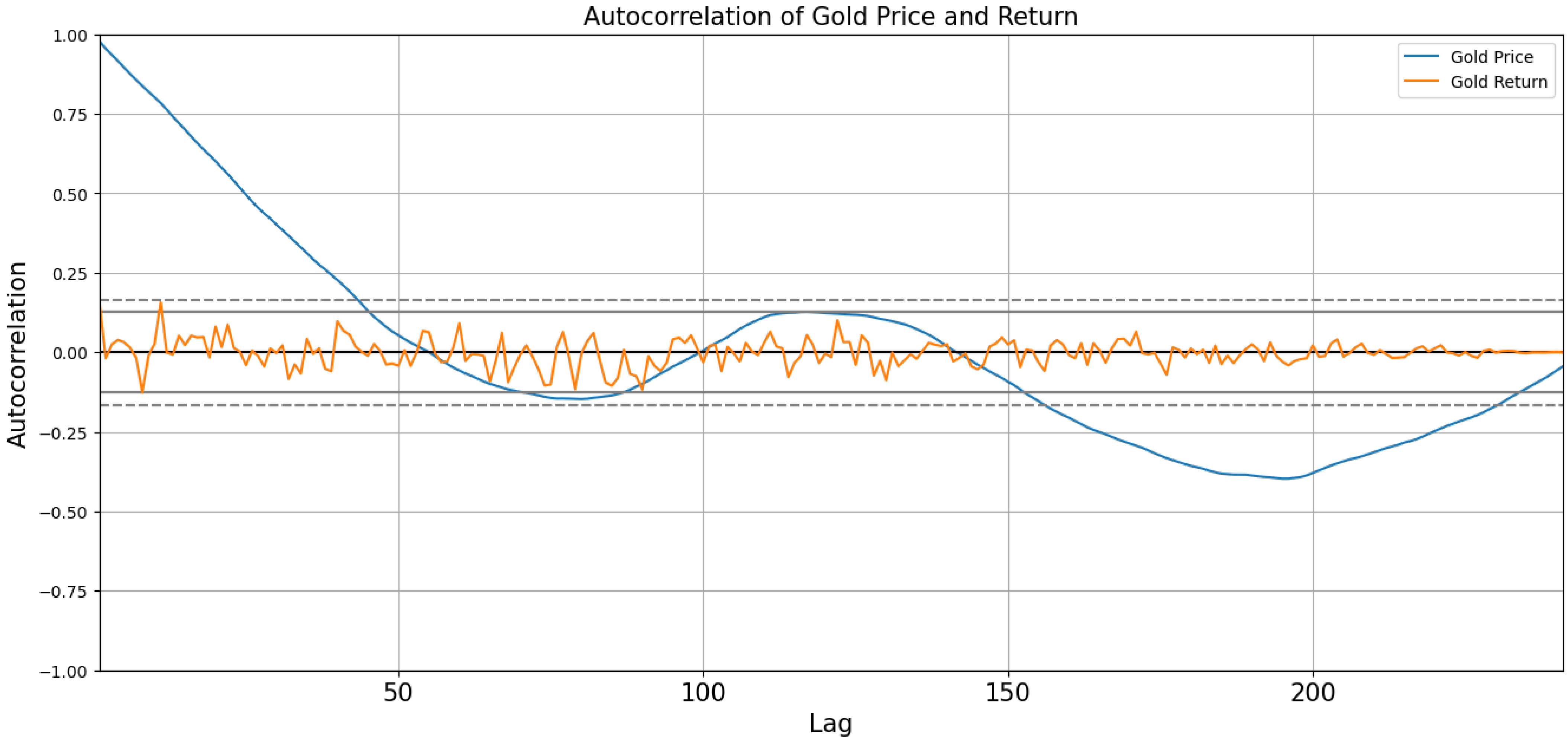


PRICE CORRELATION (TOP 15 NEGATIVE)

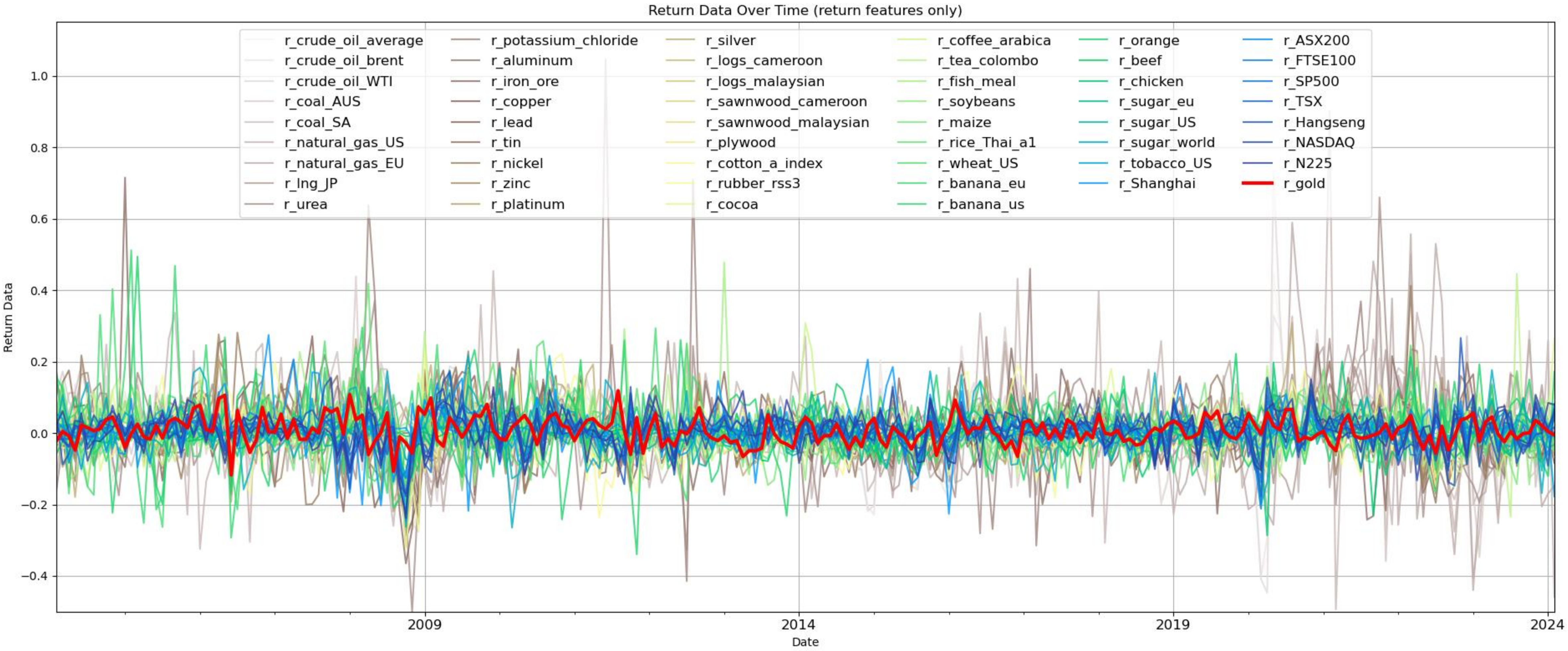


There are only 15 features negatively correlated with gold return

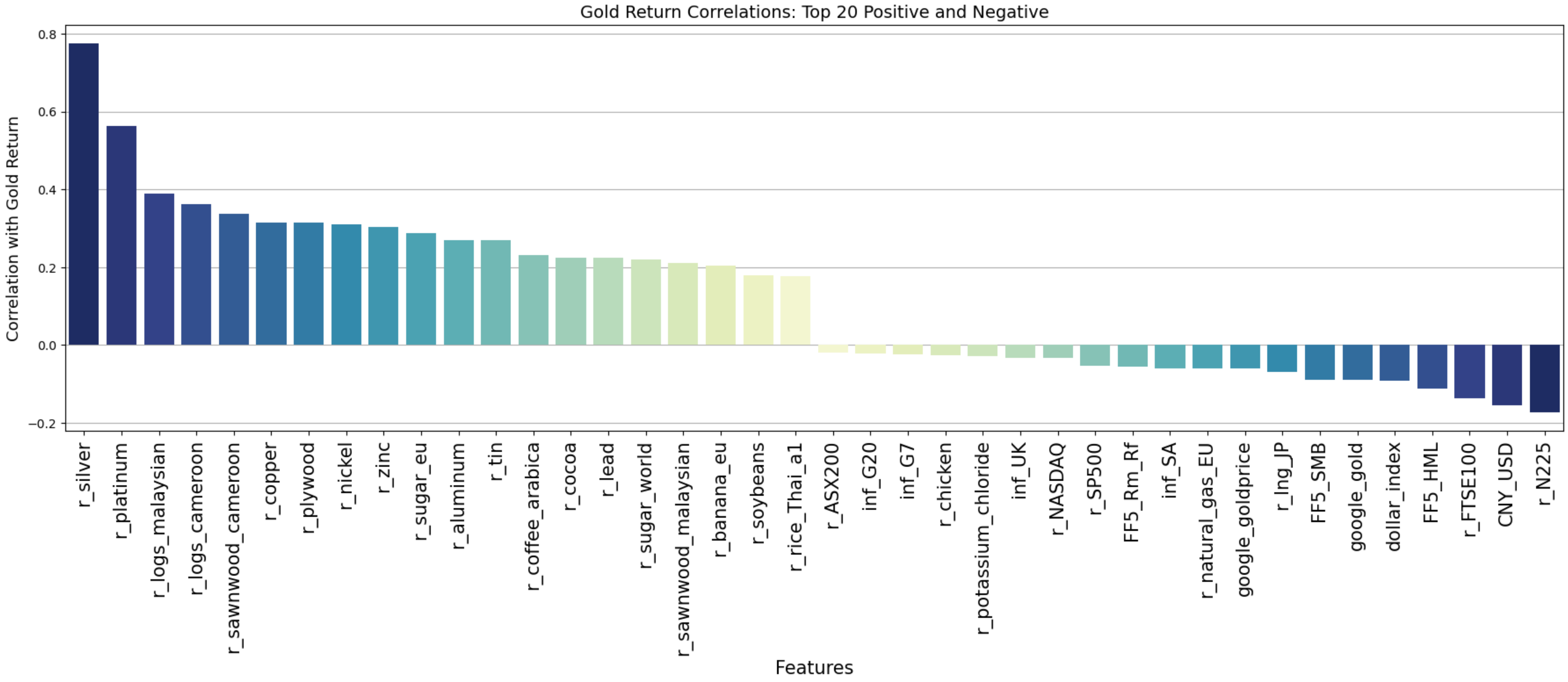
AUTOCORRELATION ANALYSIS



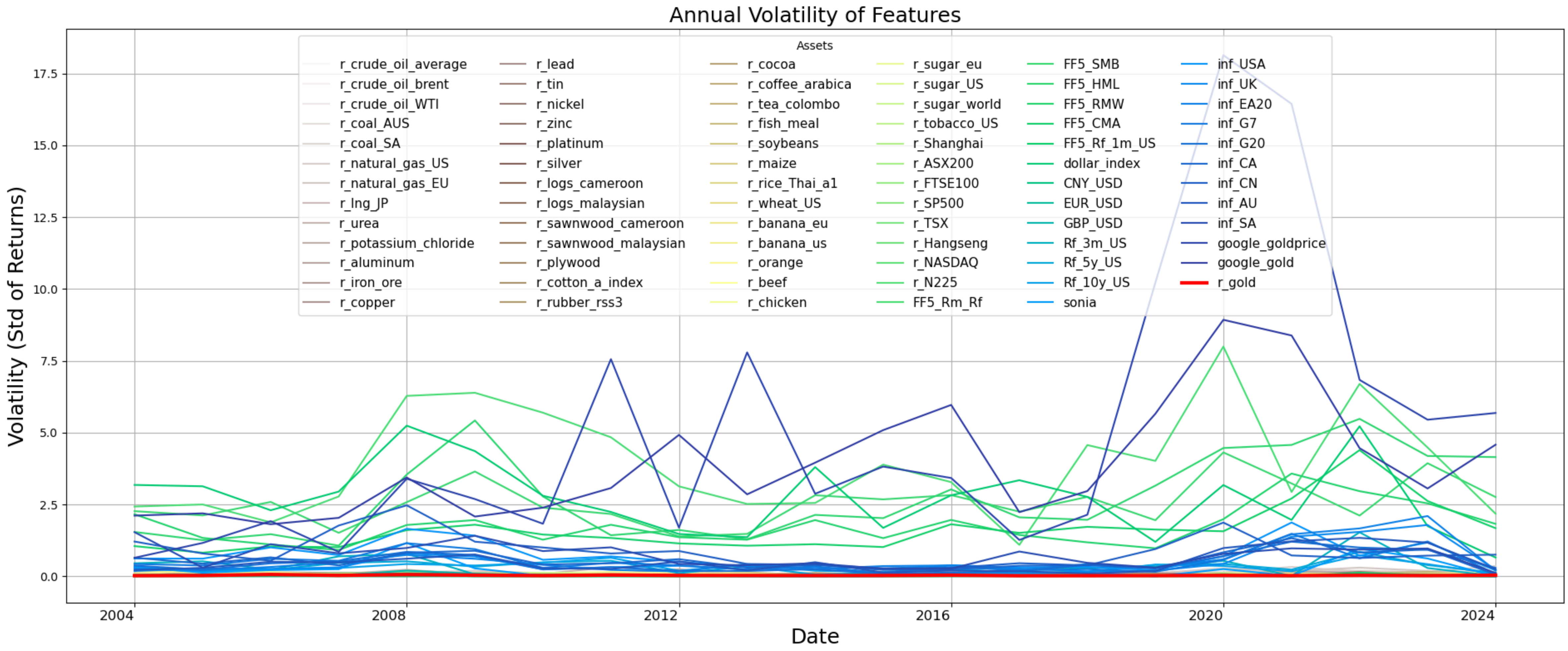
RETURN DATA (RETURN FEATURES ONLY)



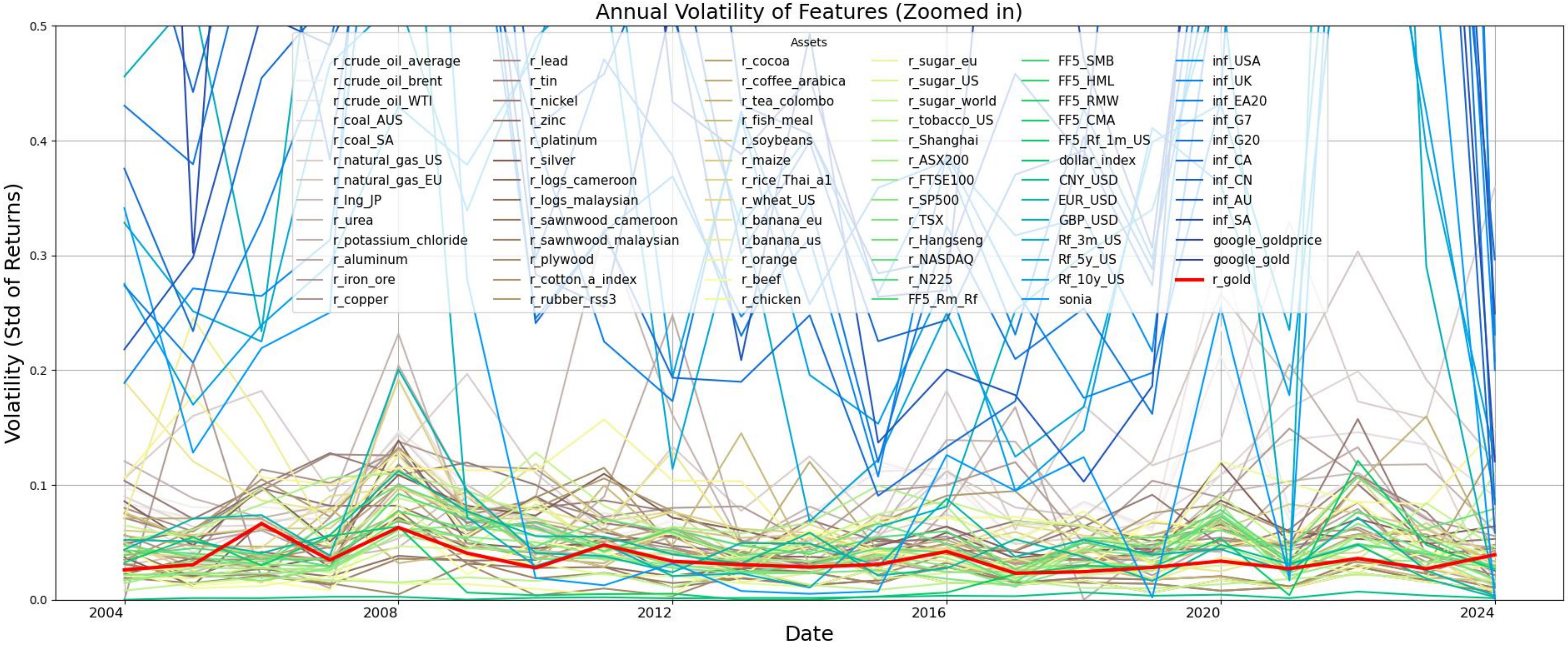
RETURN CORRELATIONS: TOP 20 POSITIVE AND NEGATIVE



ANNUAL VOLATILITY OF RETURN

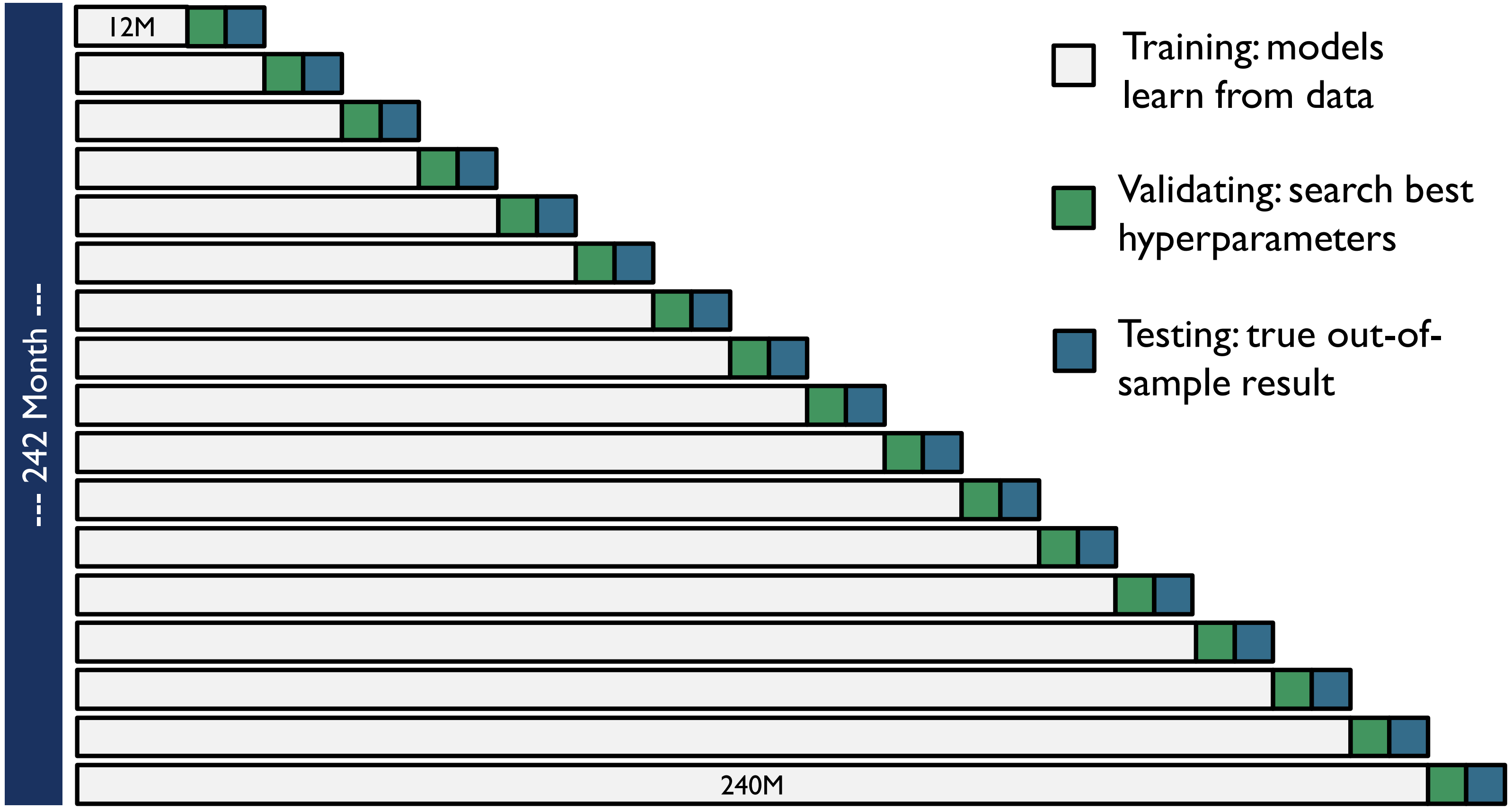


ANNUAL VOLATILITY OF RETURN (ZOOMED IN)



PREDICTION SETTINGS

Expanding Window



□ Training: models learn from data

🟩 Validating: search best hyperparameters

🟦 Testing: true out-of-sample result

Models

- Linear Models
  - Linear Regression (OLS)
  - LASSO Regression

- Tree Based Model
  - XGBoost

- Neural Networks
  - LSTM (fixed sequence)
  - LSTM (expanding sequence)

Training and Tuning

- Model refit each expansion to search best hyperparameters
- Grid Search in hyperparameters space
- Initial window 12M, total refit 228 times

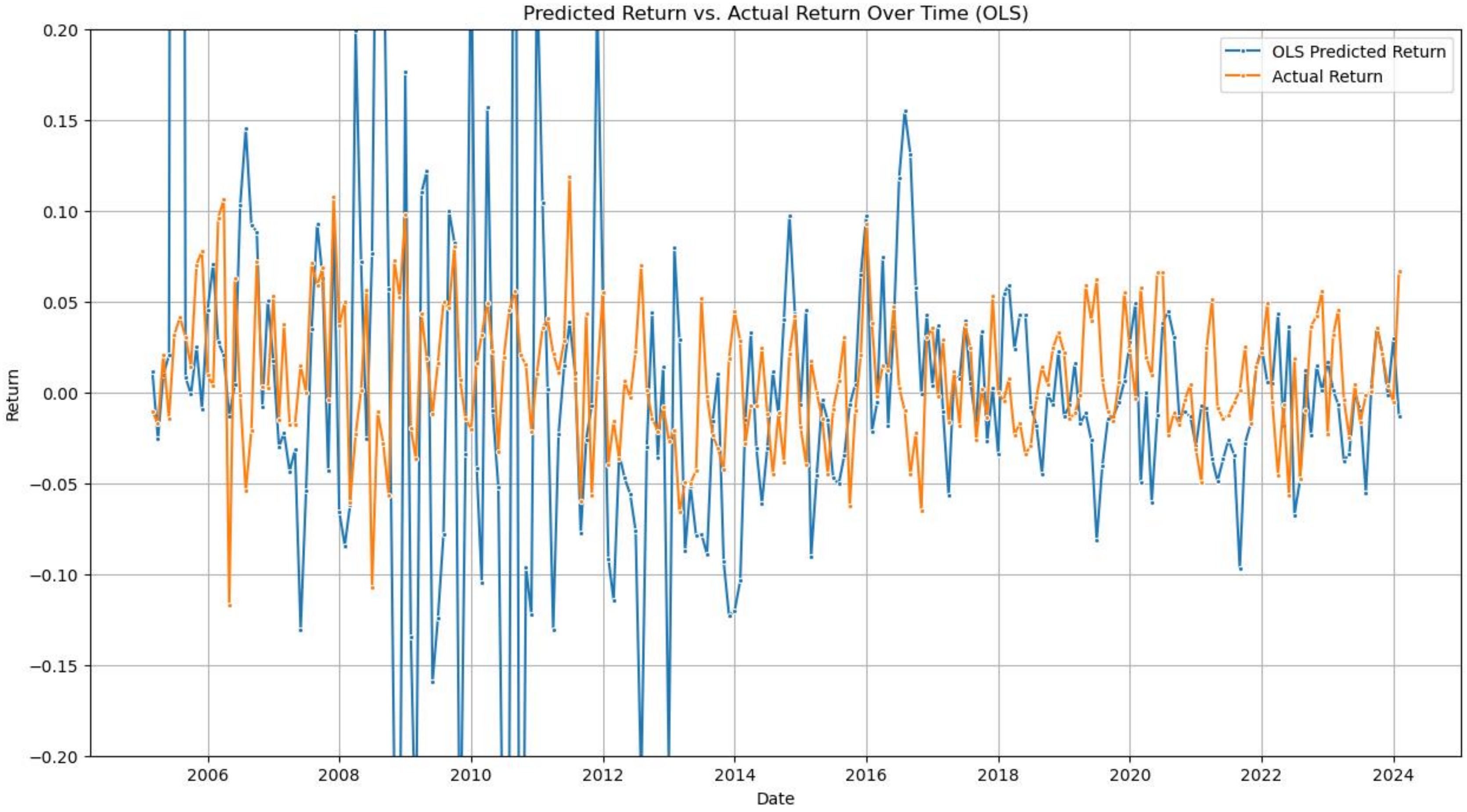
OLS RESULTS

Model performance and hyperparameters

MSE: 0.0280

Sign Acc: 55%

- No hyperparameters
- In earlier periods predictions were extremely unstable
- Suffered from multicollinearity and non-stationary data (such as Google Search Trends)



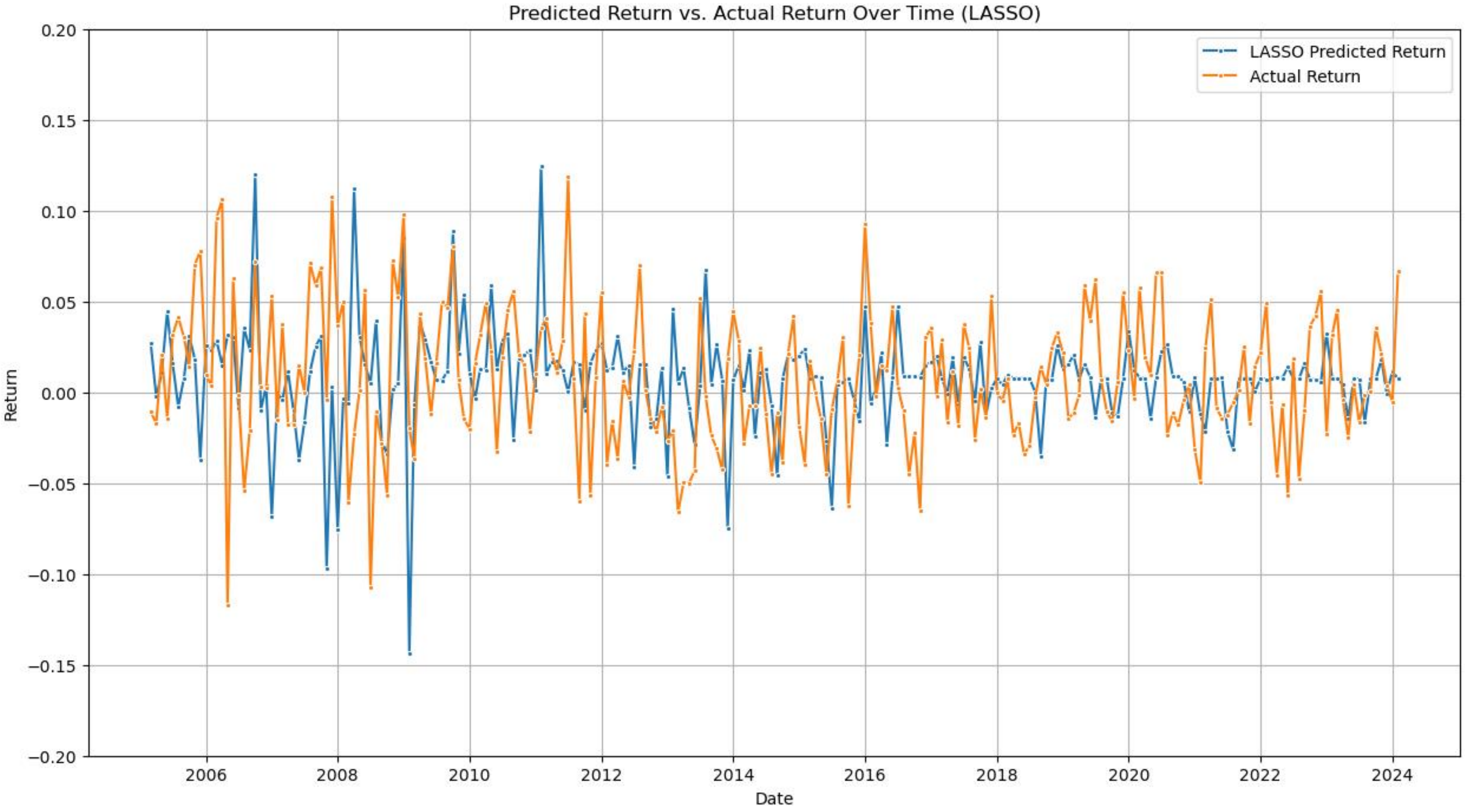
LASSO RESULTS

Model performance and hyperparameters

MSE: 0.0019

Sign Acc: 58%

- Hyperparameters searched
  - Lambda
- Predicted return movement and some sign switching before 2018
- Fail to identify sign switching in most periods after 2018



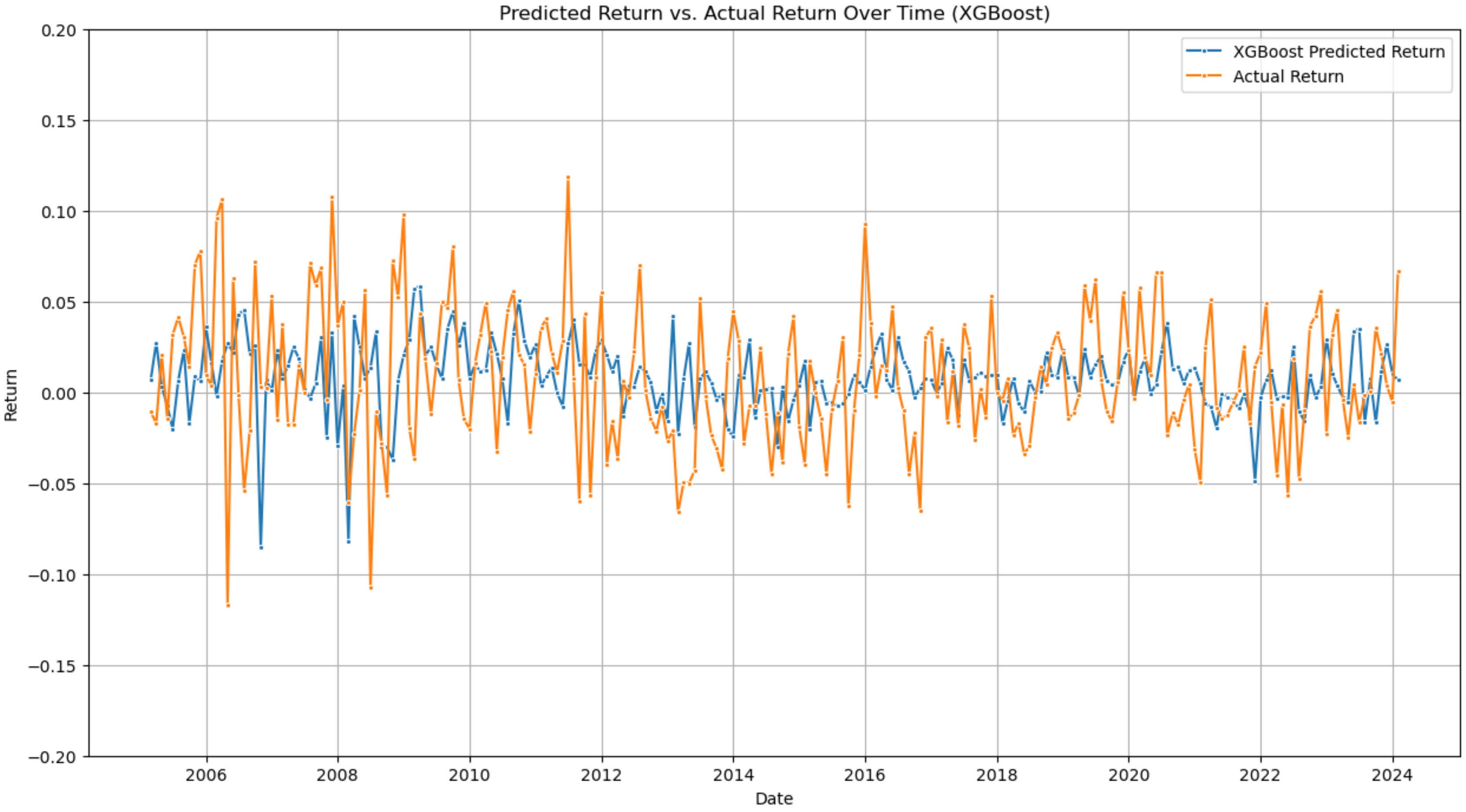
XGBOOST RESULTS

Model performance and hyperparameters

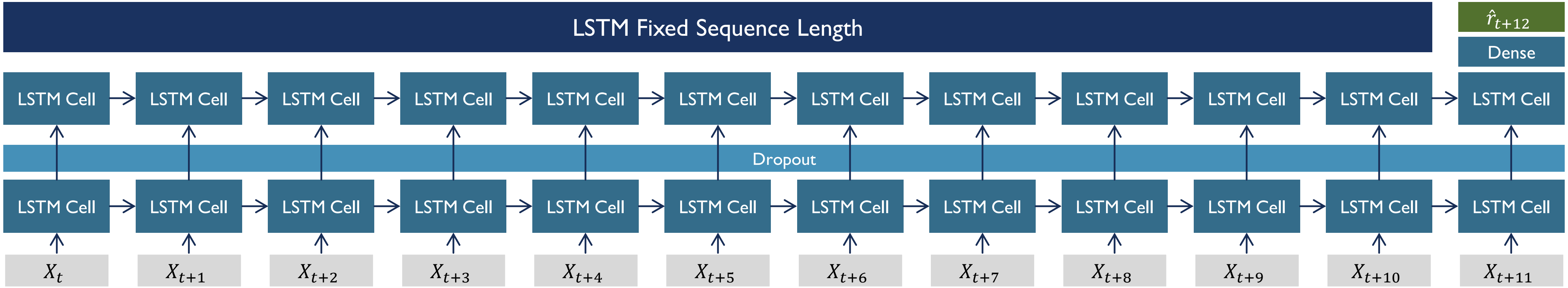
MSE: 0.0017

Sign Acc: 59%

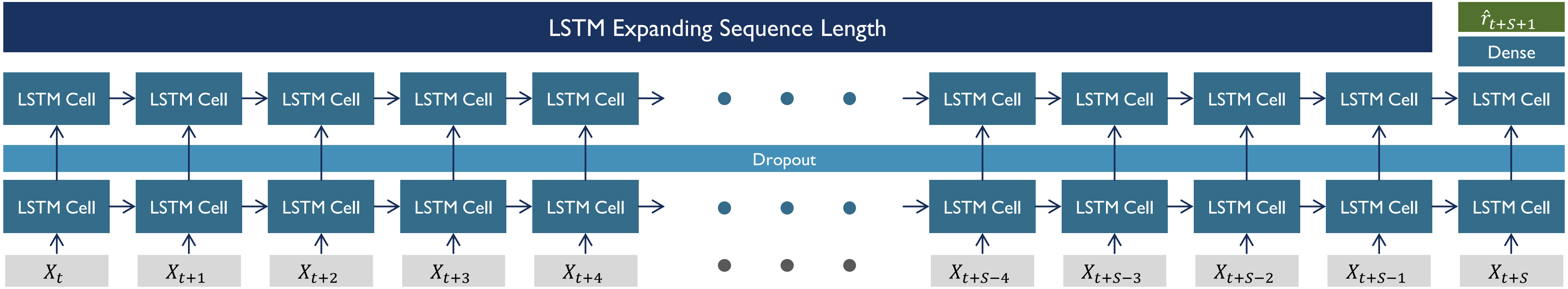
- Hyperparameters searched
  - Learning rate
  - Max depth
  - Lambda
  - Fraction of feature sample
  - Number of boosted tree
  - Subsample
- Predicted movement and sign switching
- Stable prediction - no extreme predicted return
- Lowest MSE among all models



LONG SHORT-TERM MEMORY (LSTM)



LSTM Fixed Sequence Length



LSTM Expanding Sequence Length

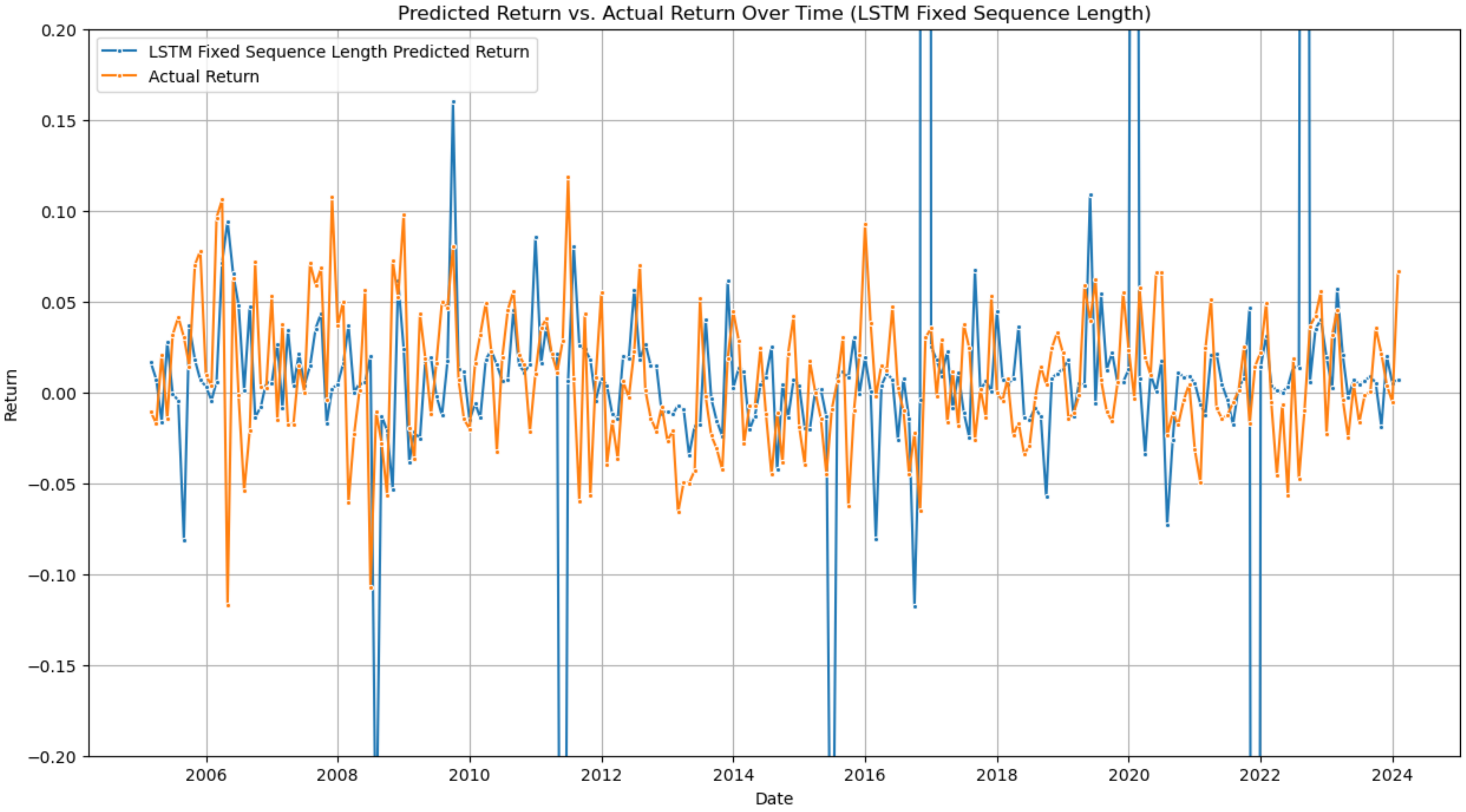
LSTM FIXED SEQUENCE LENGTH

Model performance and hyperparameters

MSE: 0.1958

Sign Acc: 59%

- Hyperparameters searched
  - Number of layers
  - Hidden dimension size
  - Layer dropout rate
  - LSTM bias
  - Weight decay (Lambda)
  - Learning rate
- Predicted return movement and sign switching
- Extreme predicted return value exist
- Highest MSE among all models



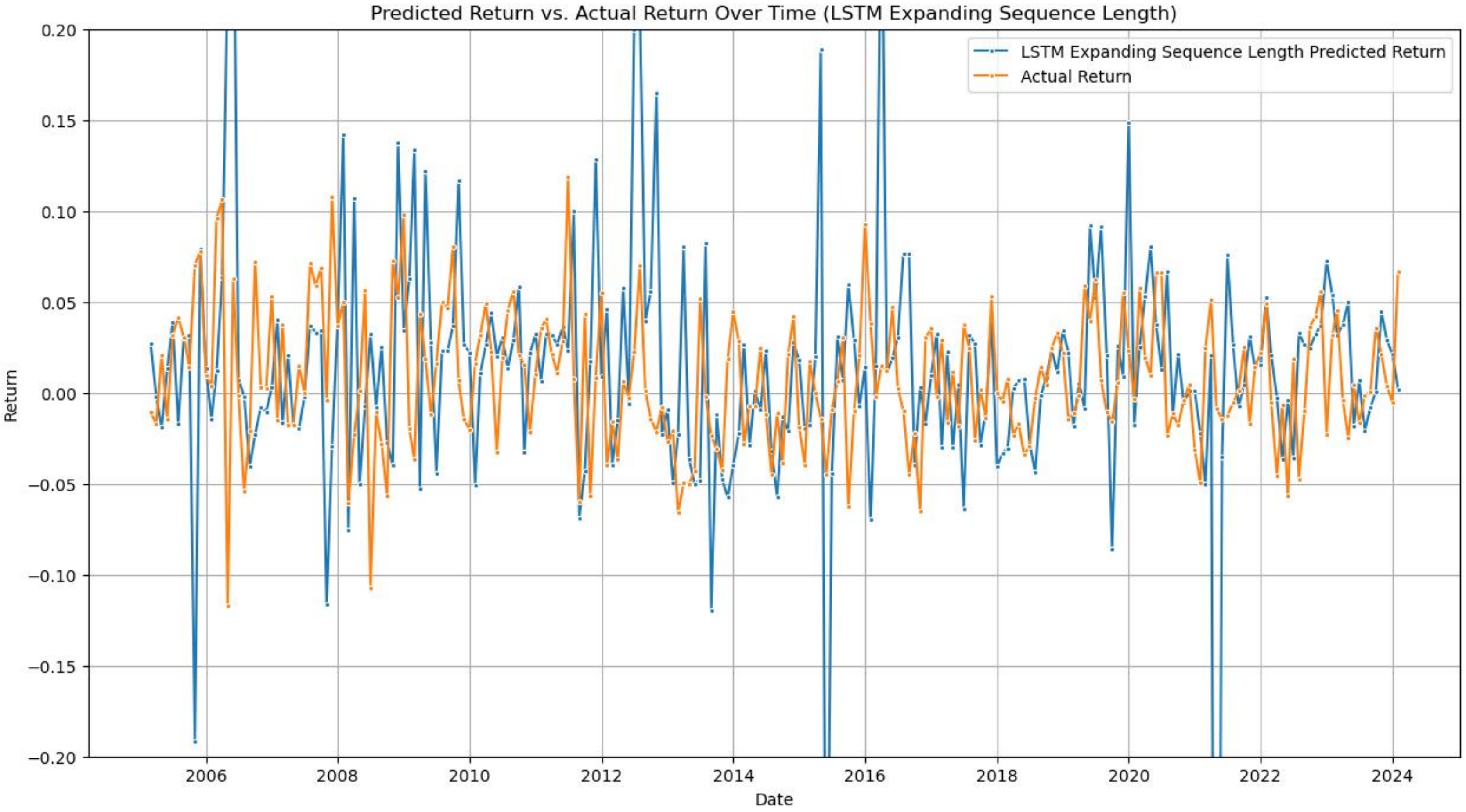
LSTM EXPANDING SEQUENCE LENGTH

Model performance and hyperparameters

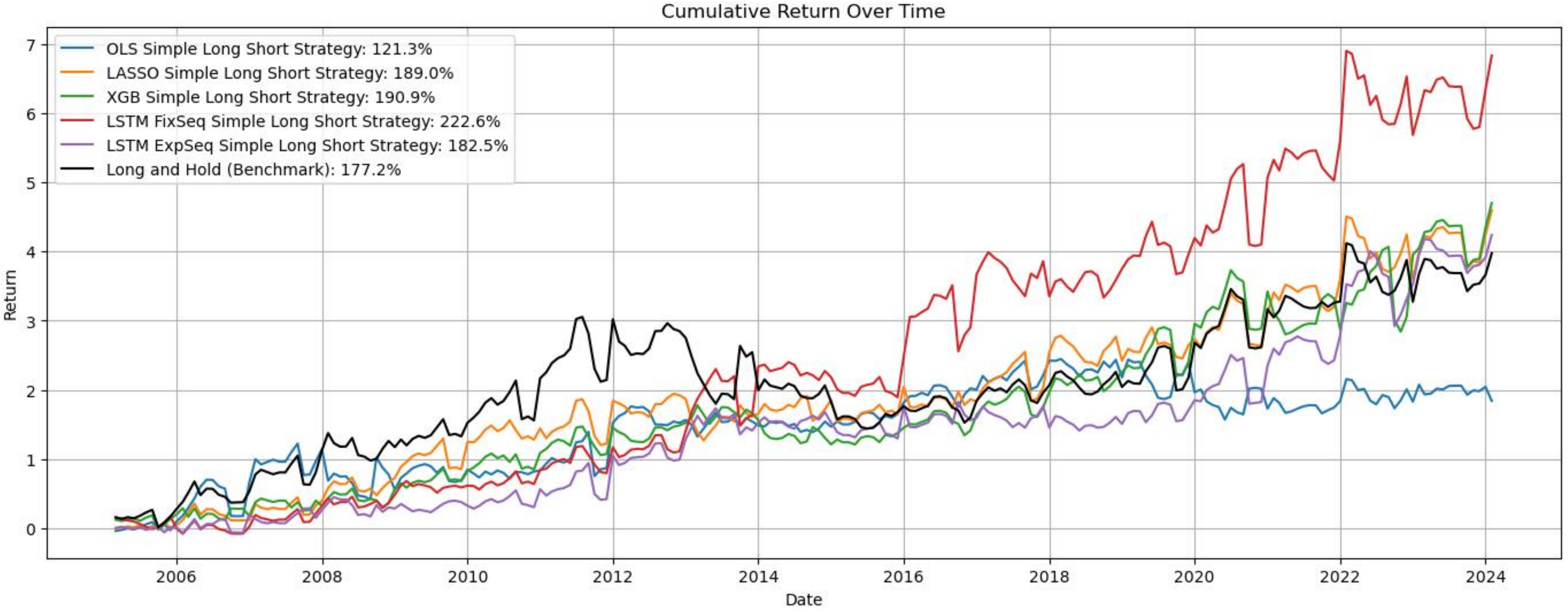
MSE: 0.0072

Sign Acc: 57%

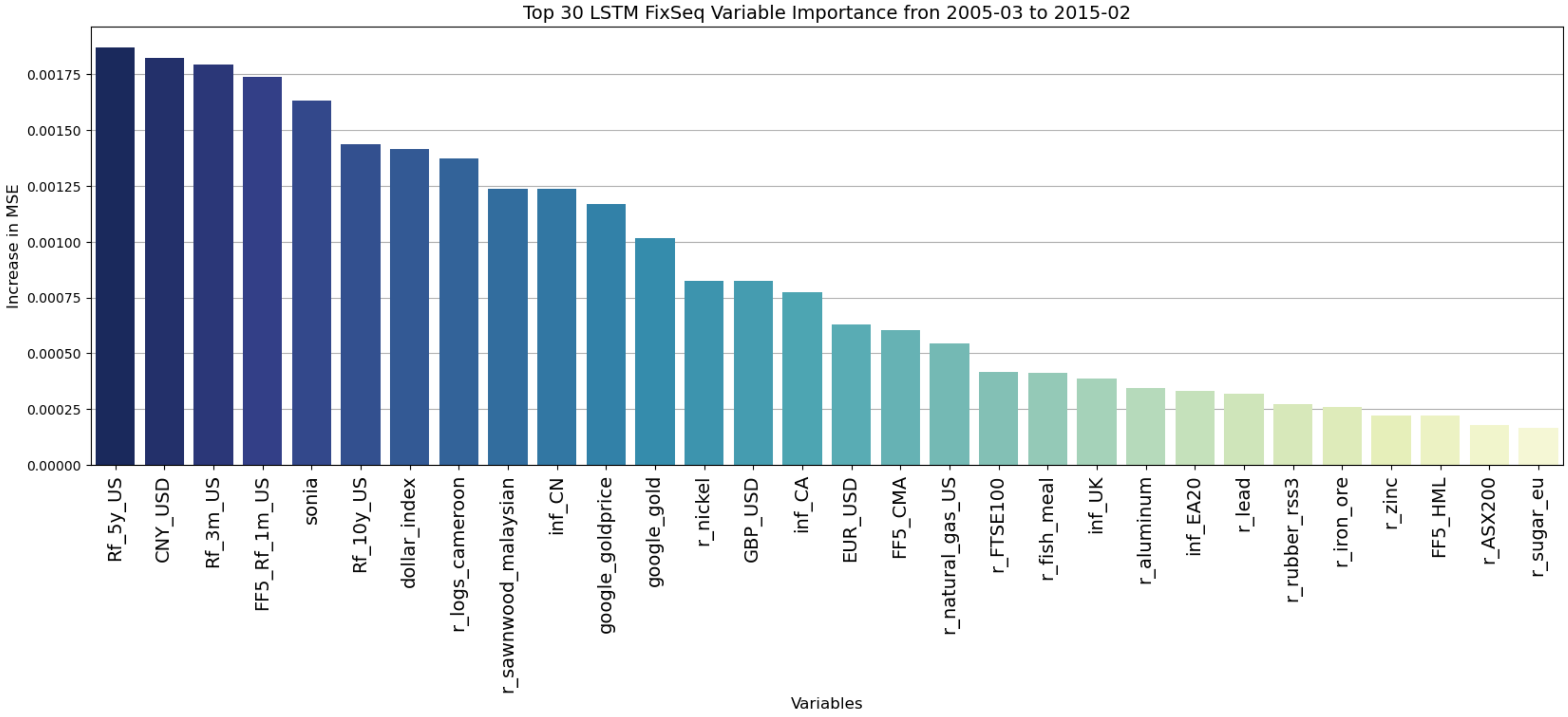
- Hyperparameters searched
  - Number of layers
  - Hidden dimension size
  - Layer dropout rate
  - LSTM bias
  - Weight decay (Lambda)
  - Learning rate
- Predicted return movement and sign switching
- Extreme predicted return value exist



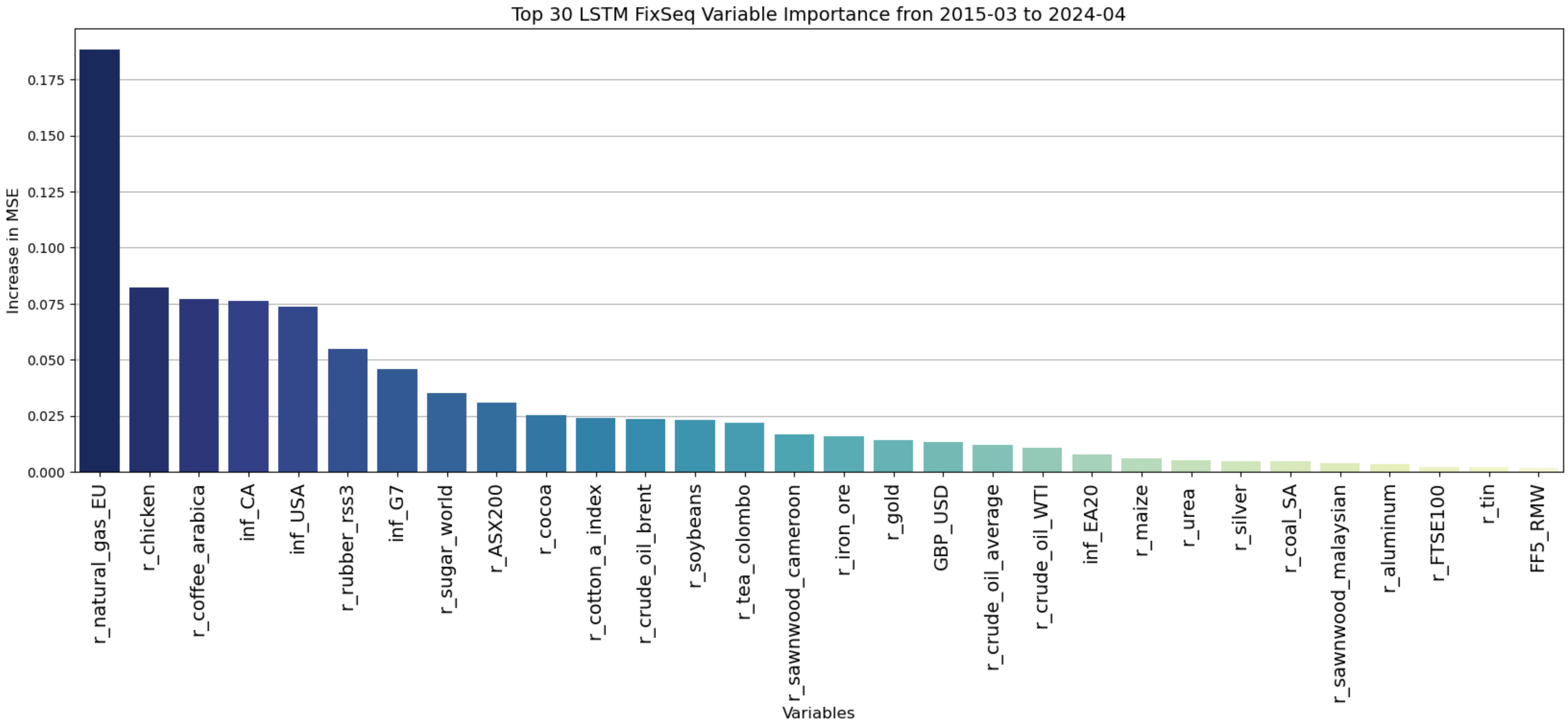
SIMPLE LONG-SHORT STRATEGY CUMULATIVE RETURN USING PREDICTION



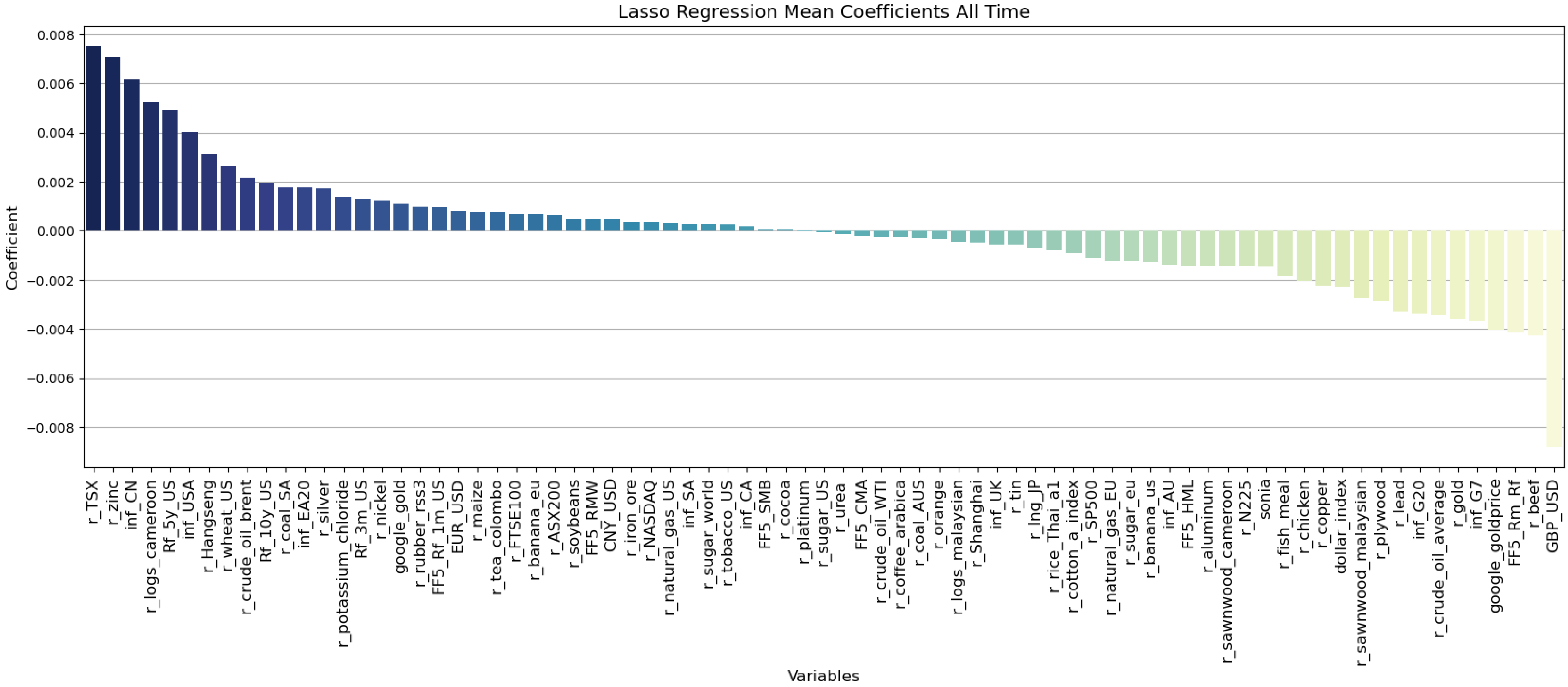
LSTM VARIABLE IMPORTANCE (2005-2015)



LSTM VARIABLE IMPORTANCE (2015-2024)

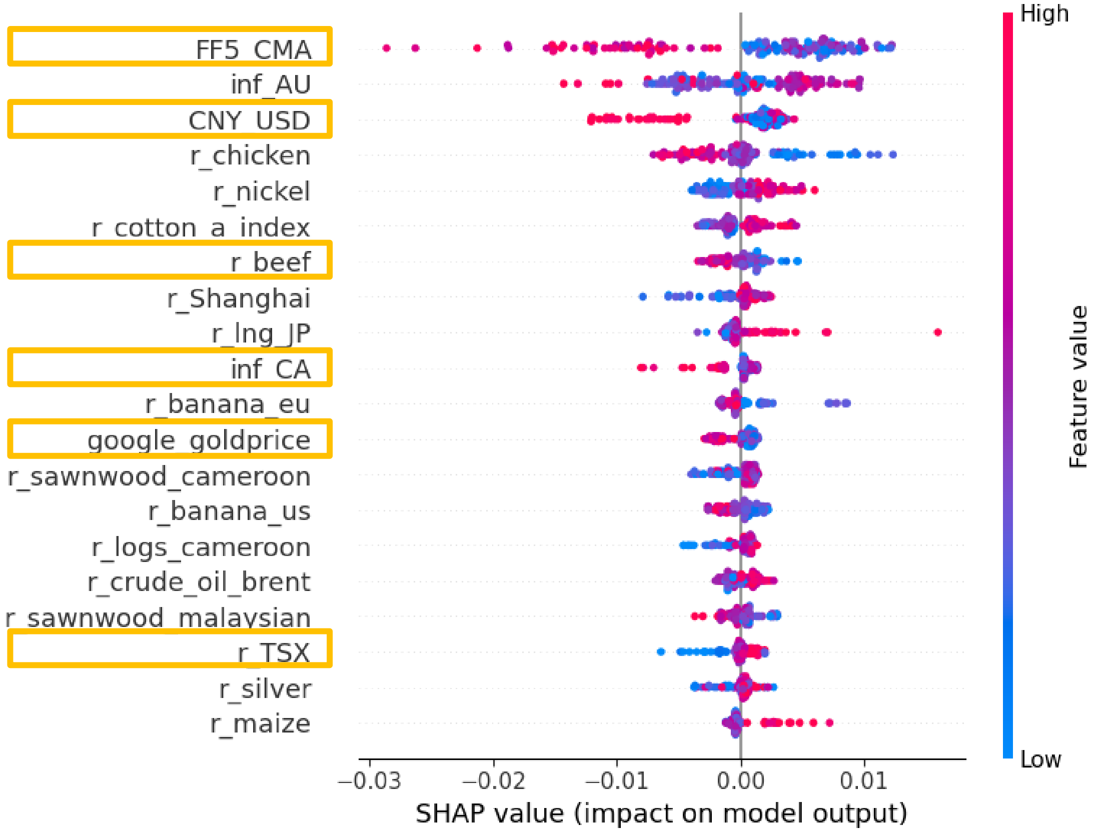


LASSO REGRESSION MEAN COEFFICIENTS ALL TIME

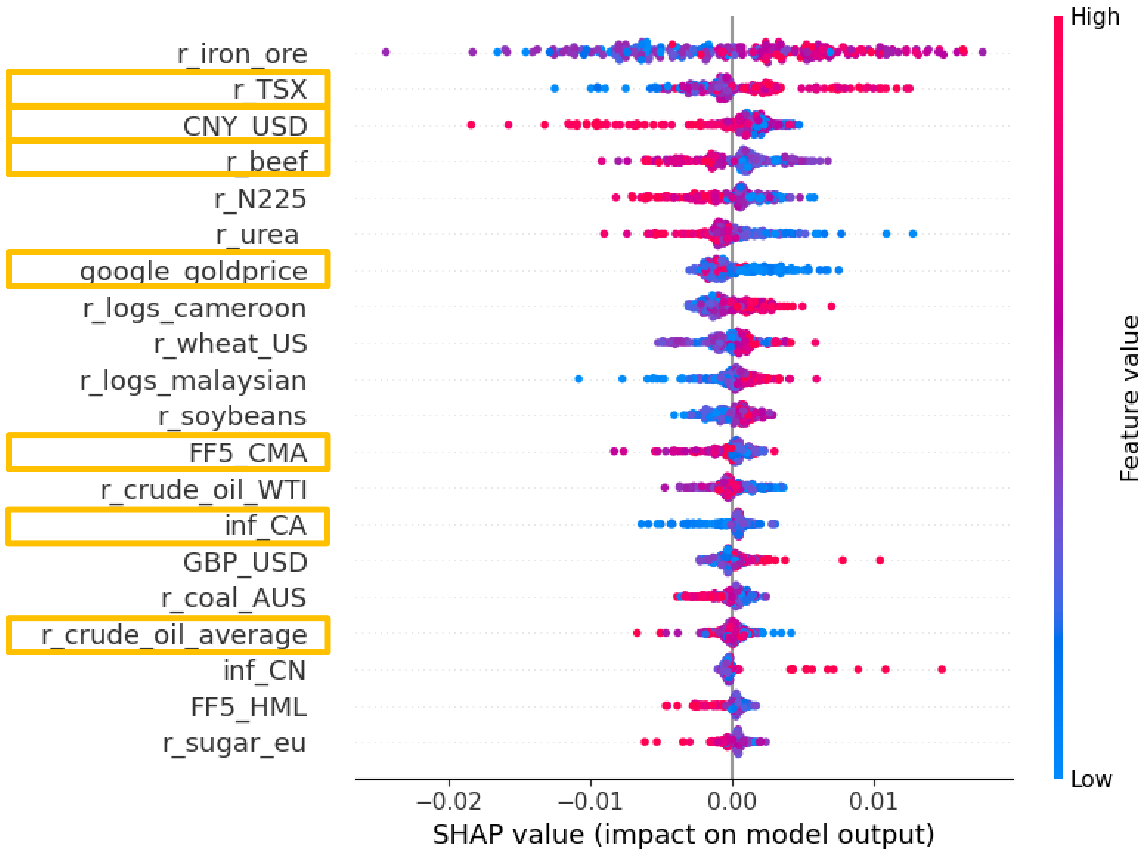


XGBOOST SHAPLEY VALUE (SAMPLES)

Feb-2015



Feb-2024



CONCLUSION

Project Limitations

- **Computational constraints**: search space was not wide enough; some model might perform better when search space is large such as XGBoost
- **Limited data**: only 20 years of observations, some data only available in quarterly such as inflation for some countries
- **No ensemble**
- **Interpretation bias**

Data Source

- Inflation of major countries: OECD Consumer price indices (CPIs, HICPs), COICOP 1999
- Commodity prices: The World Bank Pink Sheet Monthly
- US Risk Free Rate 3M, 5Y and 10Y: Federal Reserve Bank of St. Louis, FRED
- SONIA Daily: Bank of England
- Fama and French 5: Kenneth R. French
- Stock, currency and index: Yahoo Finance
- Google Trends: Google

Reference

- Dhanush, N. et al., 2021. Prediction of Gold Price using Deep Learning. IEEE R10-HTC, pp. 1-5.
- Sami, I. and Junejo, K.N., 2017. Predicting Future Gold Rates using Machine Learning Approach. *IJACSA*, 8(12), pp. 1-8
- Cohen, G. and Aiche, A., 2023. Forecasting gold price using machine learning methodologies. *Chaos, Solitons & Fractals*, 175, 114079

THANK YOU FOR LISTENING!

Q&A