Commodities Time Series Prediction using Relevance Filtering and CEnet

Andong Zhou (az1418), Anurag Rathore (ar6634),
Jiayu Shi (js11603)

NYU Capstone Project
hosted by
Bank of America

Introduction

A commodity is a basic good that can be exchanged, bought, or sold. Common examples of commodities include grains, gold, beef, oil, and natural gas. The need for financial markets in commodities appeared when producers and consumers had to hedge the market price risk for their produce at the time of harvest or production. For investors, commodities are used to diversify their portfolios or hedge against inflation and times of stock market volatility.

Due to such importance of commodities, scholars are curious about the price prediction with time series of commodities. However, well-known traditional approaches have a few shortcomings in predicting the prices of commodities. They are insufficient in processing observations and feature selection. In order to improve out-of-sample prediction efficiency, we introduce relevance filtering (Czasonis et al., 2021) and C-Enet (Rapach and Zhou, 2020) to build a sparse matrix before modeling and then train the data. As a result, we succefully reduce the prediction error compared to benchmark for both oil group and metal group.

Motivation

As most commonly known approaches when tackling time series forecasting problems, recursive models (expanding window) and rolling models (fixed-length window) are often used to achieve predictions for a series. While such approaches may seem intuitive since they incorporate the most recent observations for every prediction, they could not take into account the relevance between the data point being predicted and the observations being fitted, i.e. neglecting the fact that the pattern of the observations itself is time-varying.

Furthermore, at every prediction, features are often included in whole into a single multi-factor model despite feature selection approaches may also be considered. For example, LASSO regression does feature selection naturally. In linear settings, the disadvantages of multi-variate models are widely discussed and such shortcomings would prevail for mishandled feature sets. Yet, we have come to a more straight-forward reasoning of why multi-factor models may be inefficient, i.e., for each historical observations, some features might show relevance to the current but others might not.

We intend to improve out-of-sample prediction efficiency by filtering observations on both the feature dimension and time dimension. To be more specific, a sparse matrix would be extracted from an observation-feature matrix and only the non-sparse elements are considered relevant to the current and would be used in fitting the model. To achieve so, we apply relevance filtering, a method brought up by Czasonis et al. (2021) combined with C-Enet which is introduced by Rapach and Zhou (2020).

Data

The data used for our analysis is time-series data which we imported through 'yfinance'. We import prices of oil and primary metals. We also collect prices of stock market index, VIX, treasuries, transpot fees, commodity prices and indices, and fundamentals of commodity, which will be used as independent variables. The dates range from 01/01/2016 to 04/30/2022. The commodities we are trying to predict in our study are oil, gold, silver and copper. To be more specific, we use both Brent Oil and WTI Oil price as the indicators for oil futures prices, and we refer to COMEX prices for gold, silver and copper. Fig.1 showcases the standardized prices of the 5 commodities of interest. As can be seen, WTI oil price went negative during 2020 Q1 and we will take that into account and filter such outliers alike before we continue with our predictive research. To further validate our argument on the choice of commodities, we generated a correlation matrix (Fig.2) for all these commodities and observed a strong correlation among them. During the time window of our analysis, besides the correlation between the two oil series and gold and silver, we found copper and silver to have the highest correlation of 0.73 followed by WTI oil and copper. We will further analyze the time series by breaking it into subsets.

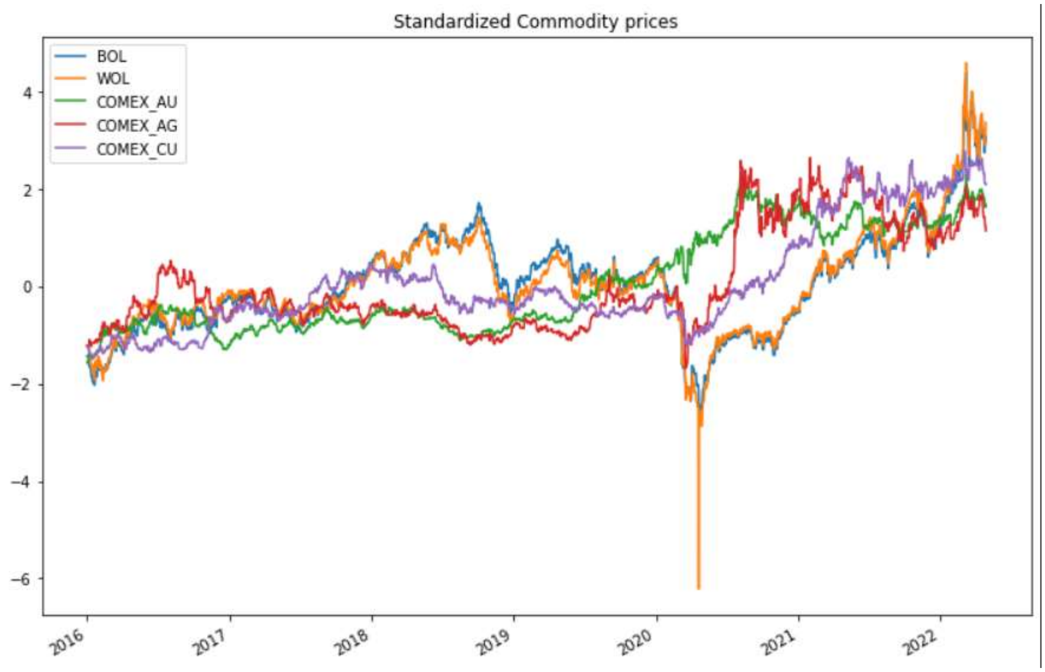
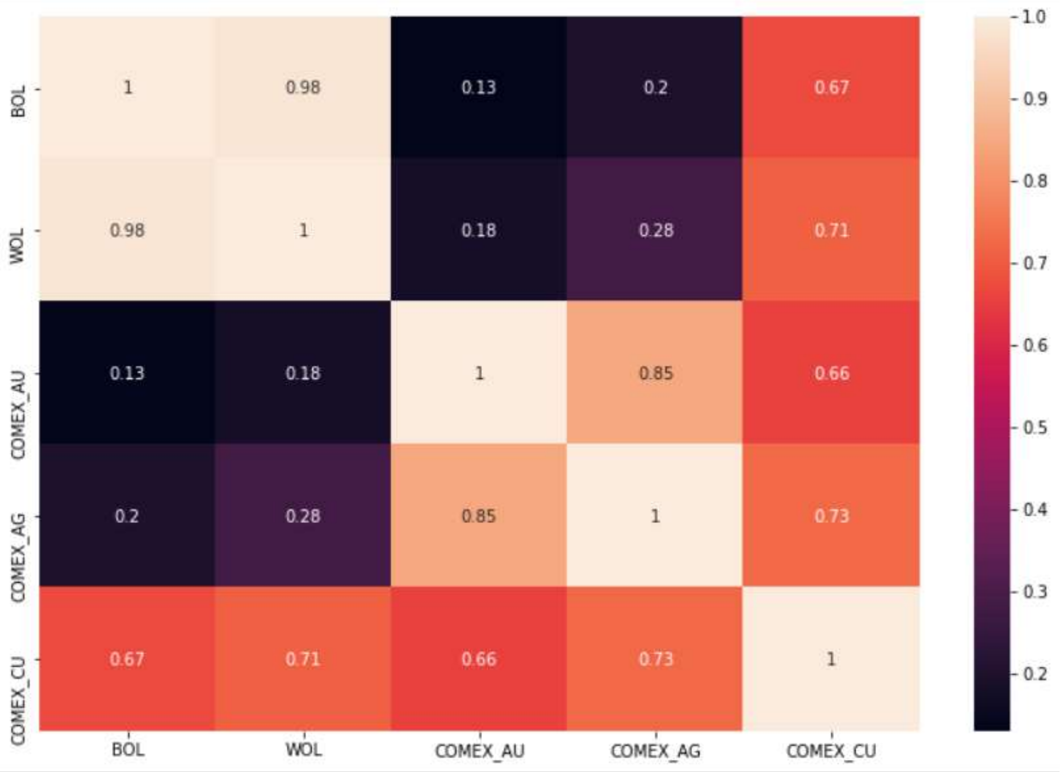


Figure 1. Standardized Commodity Prices



	BOL	WOL	COMEX_AU	COMEX_AG	COMEX_CU
BOL	1	0.98	0.13	0.2	0.67
WOL	0.98	1	0.18	0.28	0.71
COMEX_AU	0.13	0.18	1	0.85	0.66
COMEX_AG	0.2	0.28	0.85	1	0.73
COMEX_CU	0.67	0.71	0.66	0.73	1

Figure 2. Commodity Correlation Matrix

Methodologies

1 Relevance Filtering

We adopted the method introduced by Czasonis et al., 2021).

For any pair of observations x_i and x_j, each of which is a row vector of values for a set of independent variables X, we define similarity and informativeness in terms of them and Ω^{-1}, the inverse of the full sample covariance matrix of X, as shown in Equations (1) and (2).

$$sim_{ij} = sim(x_i, x_j) = -\frac{1}{2}(x_i - x_j)\Omega^{-1}(x_i - x_j)' \tag{1}$$

$$info_i = info(x_i) = \frac{1}{2}(x_i - \overline{x})\Omega^{-1}(x_i - \overline{x})' \tag{2}$$

Similarity is the squared Mahalanobis distance between x_i and x_j, multiplied by $-\frac{1}{2}$. The negative sign converts the notion of distance into one of closeness (similarity). The factor of $\frac{1}{2}$ offsets the double counting that occurs from the identical multiplication of x_i with x_j and x_j with x_i.

Informativeness is the squared Mahalanobis distance between x_i and $\overline{x}$, multiplied by $\frac{1}{2}$. Similarly, the function of factor of $\frac{1}{2}$ is to offset the double counting that occurs from squaring $(x_i - \overline{x})$. The positive value is retained here because the point of interest is how dissimilar or distant the observations are from the average values.

Relevance is defined as in Equation (3).

$$r_{ij} = r(x_i, x_j) = sim_{ij} + info_i + info_j \tag{3}$$

To generalize the calculation of relevance, we have

$$r_{it} = -\frac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' + \frac{1}{2}(x_i - \overline{x})\Omega^{-1}(x_i - \overline{x})' + \frac{1}{2}(x_t - \overline{x})\Omega^{-1}(x_t - \overline{x})' \tag{4}$$

$$r_{it} = x_t\Omega^{-1}x_i{}' - x_t\Omega^{-1}\overline{x}' - \overline{x}\Omega^{-1}x_i{}' + \overline{x}\Omega^{-1}\overline{x}' \tag{5}$$

$$r_{it} = (x_i - \overline{x})\Omega^{-1}(x_i - \overline{x})' \tag{6}$$

Within our context of time series prediction, for each prediction we make, we applied above relevance calculation for each observation and then derived a partial-sample regression with only the observations we consider relevant, i.e. $r_{it} > 0$.

2 *Combination Elastic Net*

The notion and concept of Combination Elastic Net (abbreviate for C-Enet) is introduced by Rapach and Zhou (2020).

Conventional multiple predictive regression takes the form

$$r_t = \alpha + \sum_{j=1}^{J} \beta_j x_{j,t-1} + \varepsilon_t \tag{7}$$

Equation (7) could be straightforwardly apply to generate an out-of-sample forecast of r_{t+1} based on $x_{j,t}$ for $j = 1, \dots, J$ and data available through t:

$$\hat{r}_{t+1|t}^{OLS} = \hat{\alpha}_{1:t}^{OLS} + \sum_{j=1}^{J} \hat{\beta}_{j,1:t}^{OLS} x_{j,t} \tag{8}$$

Rapach and Zhou (2020) considered a combination forecast that takes the form of a simple average of the univariate predictive regression forecasts based on $x_{j,t}$ for $j = 1, \dots, J$ in Equation (7)

$$\hat{r}_{t+1|t}^{C} = \frac{1}{J} \sum_{j=1}^{J} \hat{r}_{t+1|t}^{(j)} \tag{9}$$

They showed that, in contrast to the conventional multiple predictive regression forecast in Equations (7), the combination forecast in Equation (9) is able to deliver out-of-sample accuracy gained on a much more consistent basis over time. However, a potential drawback to Equation (9) is that it may “overshrink” the forecast to the prevailing mean, thereby neglecting substantive relevant information in the predictor variables. In an effort to improve the combination forecast by exploiting more of the relevant information in the predictor variables (while still avoiding overfitting), Rapach and Zhou (2020) then considered the Granger and Ramanathan regression:

$$r_t = \eta + \sum_{j=1}^{J} \theta_j \hat{r}_{t|t-1}^{(j)} + \varepsilon_t \tag{10}$$

which they estimated via the elastic net to select the most relevant univariate forecasts to include in the combination forecast.

Here we provide an alternative approach to Rapach and Zhou's (2020) C-Enet. Instead of computing univariate predictive regressions before aggregating their respective forecasts via elastic net, our predictor variables are pre-grouped into different categories and respective ***relevance-filtered*** multivariate predictive regressions are estimated for each category. Out-of-sample forecasts from different categories are then aggregated with the same means as in C-Enet.

Specific procedures can be described as follows:

Step 1 For each predictor variable group, we compute rolling partial-sample multivariate predictive regression forecasts based on Equation (7) with relevance-filtered observations and L2 norm over the holdout out-of-sample period

$$\hat{r}_{s|s-1}^{(c)} = \hat{\alpha}_{1:s-1}^{c} + \sum_{j=1}^{J_c} \hat{\beta}_{1:s-1}^{(c,j)} x_{c,j,s-1}\,,$$
$$for\ c = 1, \dots, C, s = t_1 + 1, \dots, t \text{ and } j = 1, \dots, J_c \tag{11}$$

Step 2 We estimate the Granger and Ramanathan regression in Equation (10) via the ENet over the holdout out-of-sample period:

$$r_s = \eta + \sum_{c=1}^{C} \theta_c \hat{r}_{s|s-1}^{(c)} + \varepsilon_s,$$
$$for\ s = t_1 + 1, \dots, t \tag{12}$$

Let $C_t \subseteq \{1, \dots, C\}$ denote the index set of categorical multivariate predictive regression forecasts selected by the ENet in Equation (12). When estimating Equation (12), we impose the restriction that $\theta_c \geq 0$ for $c = 1, \dots, C$. This imposes the economically reasonable requirement that a market excess return forecast with positively related to the realized excess return should be selected by the ENet in Equation (12).

Step 3 We compute the C-ENet forecast as

$$\hat{r}_{t+1|t}^{CEnet} = \frac{1}{|C_t|} \sum_{c \in C_t} \hat{r}_{t+1|t}^{(c)},$$

where $|C_t|$ is the cardinality of C_t, and $\hat{r}_{t+1|t}^{(c)}$ is given by Equation (8) for $c = 1, \ldots, C$.

Model and Experiments

1 Commodities Prediction

We chose the rate of daily changes of oil and primary metals as the dependent variable (Y) of the model. Detailed ticker information is listed below in Table 1.

Oil		Metal	
Symbol	**Description**	**Symbol**	**Description**
BOL	Brent Crude Oil Price	COMEX_AU	COMEX Gold Price
WOL	WTI Oil Price	COMEX_AG	COMEX Silver Price
		COMEX_CU	COMEX Copper Price

Table 1. Information of Oil and Metals

The inputs of model (independent variable, X) includes the rate of daily changes of stock market indices, Cboe volatility index (VIX), treasuries, transpot fees, commodity prices and indices, fundamentals of commodity, and simple combination of futures. The detailed information is listed in Table 2~8 below.

Stock Market Index			
Symbol	**Description**	**Symbol**	**Description**
DJI	Dow Jones Index	HIS	Hang Seng Index
IXIC	Nasdaq Composite Index	N225	Nikkei 225 Index
SPX	Standard & Poor's 500 Index	STI	Straits Times Index
DAX	Deutscher Aktien Index	A50	FTSE China A50 Index

FTSE	Financial Times Stock Exchange Index	FCHI	France 40 Index

Table 2. Information of Stock Market Index

VIX	
Symbol	**Description**
VIX	CBOE Volatility Index
Crude_Oil_ETF_VIX	CBOE Crude Oil ETF Volatility Index
Russell2000_VIX	Cboe Russell 2000 Volatility Index
DJI_VIX	Dow Jones Volatility Index
GOLD_ETF_VIX	CBOE Gold ETF Volatility Index
SPX_VIX	Standard & Poor's 500 Volatility Index
SPX_VIX3M	CBOE 3-Month Volatility Index

Table 3. Information of VIX

Treasury	
Symbol	**Description**
USTSR1Y	US 1 Year Treasury Bill nominal yield
USTSR10Y	US 10 Year Treasury Bond nominal yield
USTTR5Y	US 1 Year Treasury Bill real yield
USTTR10Y	US 10 Year Treasury Bond real yield
USDIDX	U.S. Dollar Index

Table 4. Information of Treasury

Transport Fee			
Symbol	**Description**	**Symbol**	**Description**
BCTI	Baltic Clean Tanker Index	BLDH	Baltic Dry Index
BDTI	Baltic Dirty Tanker Index	BNM	Baltic Panamax Index
BFGJ	Tianjin Bulk Freight Index	CGH	Baltic Capesize Index

Table 5. Information of Transport Fee

Commodity Price and Index			
Symbol	**Description**	**Symbol**	**Description**
CRB	CRB Spot Index	RJCRB_Price_Index	CRB Commodity Index
CRB_Food	CRB Foodstuffs Spot Index	COMEX_AG_mini	E-mini Gold Futures
CRB_Indust	CRB Raw Industrials Spot Index	COMEX_AU_mini	E-mini Copper Futures
CRB_Metal	CRB Precious Metals Spot Index		

Table 6. Information of Commodity Price and Index

COMEX_AG_Inv	COMEX_AU_Inv	COMEX_CU_Inv	iShares_ETF_close
iShares_ETF_Hold	LDNAG_SPOT	LDNAU_SPOT	NY_GAS_spot
SPDR_AUETF_Close	SPDR_ETF_Hold	LME_AL_Receipt	LME_AL_Inv
LME_AL_spot	LME_CU_Receipt	LME_CU_Inv	LME_CU_spot
LME_NI_Receipt	LME_NI_Inv	LME_NI_spot	LME_PB_Receipt
LME_PB_Inv	LME_PB_spot	LME_SN_Receipt	LME_SN_Inv
LME_SN_spot	LME_ZN_Receipt	LME_ZN_Inv	LME_ZN_spot
LME_CU_STS0to3	LME_NI_STS0to3	LME_PB_STS0to3	LME_SN_STS0to3
LME_ZN_STS0to3			

Table 7. Fundamentals of Commodities (Inventory, Receipt, Spot and Basis)

SPX_COMEX_AU_ratio	LME_CU_Receipt_LME_ZN_Receipt_logdiff
SPX_VIX_Russell2000_VIX_diff	COMEX_CU_BOL_logdiff
USTSR10Y_USTTR10Y_diff	COMEX_AU_COMEX_AU_mini_logdiff
USTSR10Y_USTSR1Y_diff	COMEX_AU_BOL_logdiff
CRB_CRB_Metal_logdiff	COMEX_AU_COMEX_AG_logdiff
CRB_Indust_CRB_Metal_logdiff	

Table 8. Simple Combination of Features

2 Model

Our proposed model contains relevance filter and CEnet approaches. We firstly derived respective relevance filtered daily predictions for each independent variable category on a 126-day (half-a-year) rolling basis; then we trained C-Enet on predictions of all groups with a rotational 3-year in-sample and 3-month holdout period to get a concatenated out-of-sample prediction aggregation.

Futher more, we have 2 benchmark models, implementing relevance filtered regression and Ridge regression for the predictions respectively. Both of these benchmark models derive daily predictions for each independent variable category on a 126-day (half-a-year) rolling basis. Predictions are aggregated through taking an average across the variable categories.

3 Metrics

To quantify the out-of-sample performance of our method, we mainly apply the

Campbell and Thompson (2008) R_{OS}^2 statistic. As standard in the literature, the benchmark models are the average return premium up to time t $(\overline{r_t})$. In this case, R_{OS}^2 statistic measures the proportional reduction in the mean squared forecast error od the predictive model $(MSFE_{PRED})$, comparing to the historical mean $(MSFE_{HM})$. The R_{OS}^2 statistic is calsulated by

$$R_{OS}^2 = 100 * \left(1 - \frac{MSFE_{PREP}}{MSFE_{HM}}\right) = 100 * \left[1 - \frac{\sum_{t=t_0}^{T-1}(r_{t+1} - \widehat{r_{t+1}})^2}{\sum_{t=t_0}^{T-1}(r_{t+1} - \overline{r_t})^2}\right]$$

We have also generated plots of the cumulative sum of squared error differences (CSSED) from our proposed model versus the benchmark, which is calculated by

$$CSSED_{it} = \sum_{\tau=1}^{\tau=t}(e_{Bmk,i\tau}^2 - e_{Pbrk,i\tau}^2)$$

Result

1 Oil Group

Table 9 displays the experienment result for commodities in the oil group. The proportional reduction in prediction error of our model is about 66.72%, the proportional reduction in relevance filtered prediction error is about 51.03%, and the proportional reduction in benchmark prediction error is about 31.4%.

Proportional reduction in prediction error	0.6672
Proportional reduction in relevance filtered prediction error	0.5103
Proportional reduction in benchmark prediction error	0.3140

Table 9. Proportional Reduction for Oil Group

As plot in Fig 3, the CSSED of the oil group increases on a consistent basis over time, and has a rapid growth around April 2020.

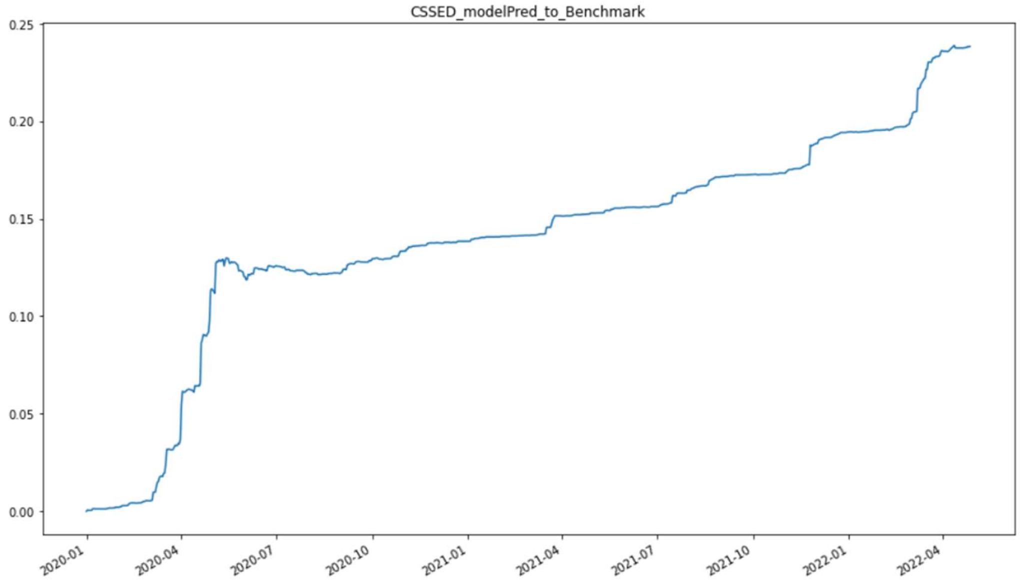


Fig 3. Oil Group Cumulative Sum of Squared Error Differences compared to Benchmark

2 Metal Group

Table 10 displays the experienment result for commodities in the metal group. The proportional reduction in prediction error of our model is about 70.15%, the proportional reduction in relevance filtered prediction error is about 48.35%, and the proportional reduction in benchmark prediction error is about 29.5%.

Proportional reduction in prediction error	0.7015
Proportional reduction in relevance filtered prediction error	0.4835
Proportional reduction in benchmark prediction error	0.2950

Table 10. Proportional Reduction for Metal Group

According to Fig 4, the CSSED of the metal group has a rapid growth in 2020 Q3, and increases steadily in the rest of time range.

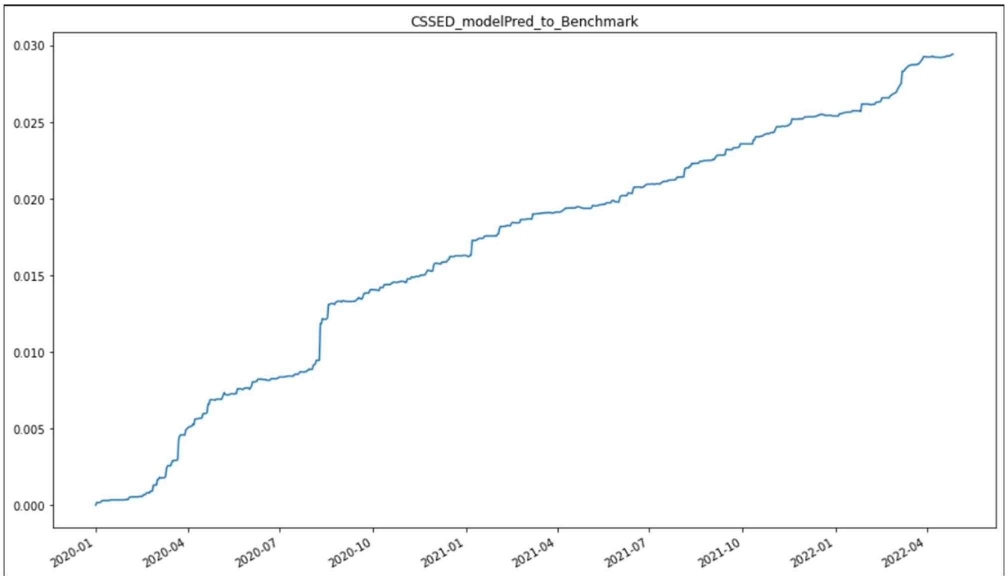


Fig 4. Metal Group Cumulative Sum of Squared Error Differences compared to Benchmark

In summary, our method outperformed conventional multivariate predictive models across all 5 commodities and 2 major commodity groups in our study in terms of the proportional reduction in prediction error and CSSED. Additionally, from the result shown in table 9 and 10, it is obvious that relevance-filtered regression alone generates significant reduction in prediction error. Additionally, C-Enet as an aggregation technique provides further power in gaining prediction accuracy on top of relevance filtering, meaning that each of the two methods are statistically proven to be inducive to accuracy gain.

Conclusion

Predicting asset prices by time series requires works, and traditional approaches may exist some flaws, such as neglecting the relevance among observations and naturally selecting features. Inspired by Czasonis (2021), Rapach and Zhou (2020), we focused on the asset class of commodity and implemented relevance filtering and Combination Elastic Net (C-Enet) in the experiments. It is evident from the results of our experiment that the proposed model is much more powerful compared with the two benchmark models we mentioned before, and reduced prediction errors in a notable proportion for both of the oil group and metal group.

References

Campbell, J. Y. & Thompson, S. B. (2008) ‘Predicting excess stock returns out of sample: Can anything beat the historical average?’, *The Review of Financial Studies*, 21(4), pp. 1509-1531, JSTOR [Online]. Available from: https://www.jstor.org/stable/40056860?seq=1(Accessed: 10 Mar 2022).

Czasonis, M., Kritzman, M. & Turkington, D. (2021) ‘Relevance’, *MIT Sloan Research Paper No. 6417-21*, SSRN [Online]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3803440 (Accessed: 13 Mar 2022).

Rapach, D. E., J. K. Strauss & G. Zhou (2010) ‘Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy’, *Review of Financial Studies*, 23(2), pp. 821–862, JSTOR [Online]. Available from: https://www.jstor.org/stable/40468327?saml_data=eyJzYW1sVG9rZW4iOiJlMDQ5ZjcyZC05MmFkLTQ5ZWYtOWYzNi1iOTM2YTQ5YTcyMDciLCJlbWFpbCI6ImpzMTE2MDNAbnl1LmVkdSIsImluc3RpdHV0aW9uSWRzIjpbImFmYmFjOTE2LTJhMTEtNDlmMC05ODc3LTMzYjM1MmJhOTk1MiJdfQ&seq=1 (Accessed: 11 Mar 2022).

Rapach, D. E. & Zhou, G. (2020). ‘Time-series and cross-sectional stock return forecasting: new machine learning methods’, *Machine Learning for Asset Management: New Developments and Financial Applications*, SSRN [Online]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3428095 (Accessed: 11 Mar 2022).