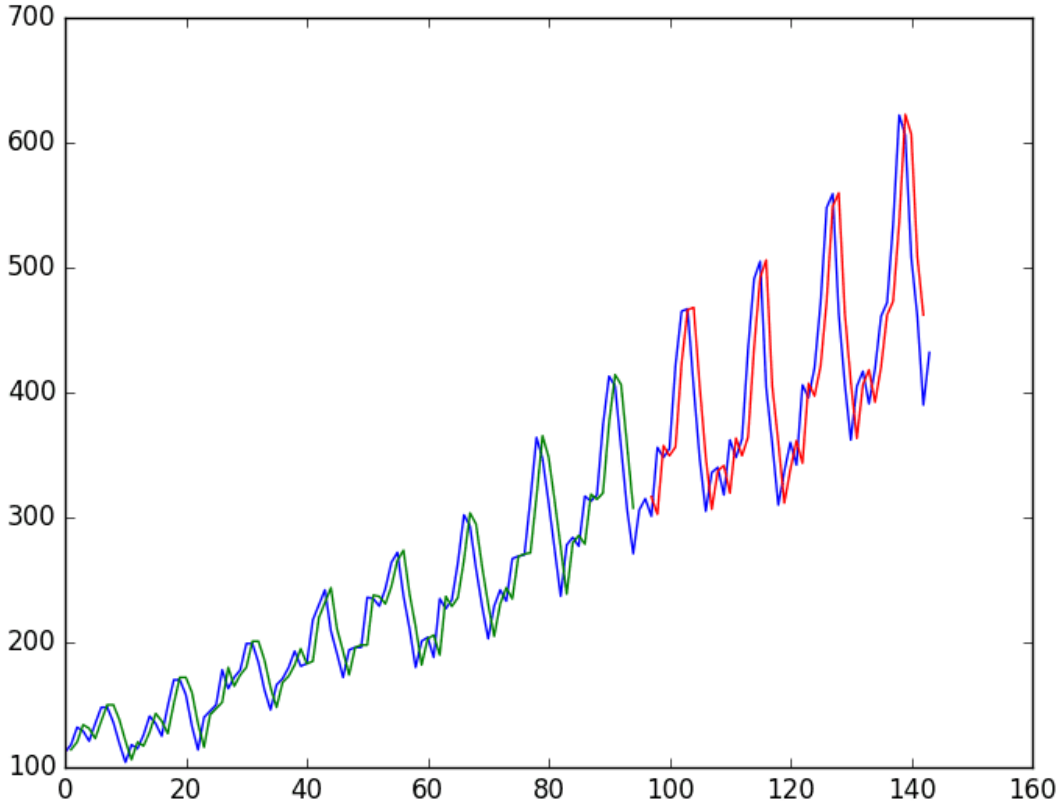
PREDICTING COMMODITY PRICES

Model Review and Selection



Team 9

July 9th 2023
MGT6203, Summer 2023

INTRODUCTION

The prediction of time series data, particularly in the context of commodity prices, has garnered significant attention due to its potential to enhance decision-making processes in various industries. This report explores the creation and testing of ML models for predicting commodity prices. By examining data preprocessing, model selection, training, and evaluation, this report aims to shed light on the efficacy and applicability of ML models in the domain of commodity price prediction. The findings and insights derived from this research can contribute to enhancing decision-making processes and optimising strategies in commodity-related industries.

HYPOTHESIS

The main hypothesis is: There is a better way to predict commodity prices better than general exponential smoothing. The most important sub-hypothesis is: Macroeconomic data and weather will be a factor in the prediction of the prices of commodities.

DATASETS USED

“Commodities prices”: The dataset comprises monthly prices of 53 commodities and 10 indexes, spanning from 1980 to 2016. The dataset is in good order, with the “Date” column set to 1st day of the month and all other columns with “double” values for prices and indices. So far, we have utilised soybean, wheat and oil prices in various models in an attempt to predict the price movements. [1]

“US weather data”: The dataset is a combination of several datasets, including average monthly temperature data of US states, covering the time period from January 1950 to August 2022. The data source for this information is the NOAA National Centers for Environmental Information. The data is again in good order, with separate “month” & “year” columns as well as other columns like “average for dataset timespan” and latitude & longitude of the geographic centres of each state.[2]

“US macroeconomic data”: The dataset consists of the basic macroeconomic features like inflation, mortgage, and unemployment rate, NASDAQ index, disposable income, personal consumption, and personal savings. It covers the time period from November

[1] On https://www.kaggle.com/datasets/vagifa/usa-commodity-prices
[2] On https://www.kaggle.com/datasets/justinrwong/average-monthly-temperature-by-us-state

1980 to May 2022. The data is from Federal Reserve Economic Data and retrieved from Kaggle.[3]

PAPERS / LITERATURE / MATERIALS

The reference landscape of ML for R is not extremely rich, or in other words, is considerably less versatile than the one existing for Python. This goes even more so for the time series niche. For example, books published can be allocated into two main categories: (i) discussing statistical/mathematical concepts on a higher level or (ii) dealing with applying the ML models for time series forecasting on an introductory level, at best. Furthermore, most books on the topic were published +5 years prior, which makes them outdated in both the coding approaches applied as well as models proposed. Nonetheless, the following work was reviewed:

- (PDF) Modelling Inflation Dynamics: A Critical Review of Recent Research (researchgate.net)[4]
- Can you predict commodity stocks using a weather forecast? by I. Colbert Medium[5]
- Climate and environmental data contribute to the prediction of grain commodity prices using deep learning[6]
- Price forecasting and evaluation: An application in agriculture. By Jon A. Brandt, David A. Bessler[7]

DATA PREPROCESSING / EDA

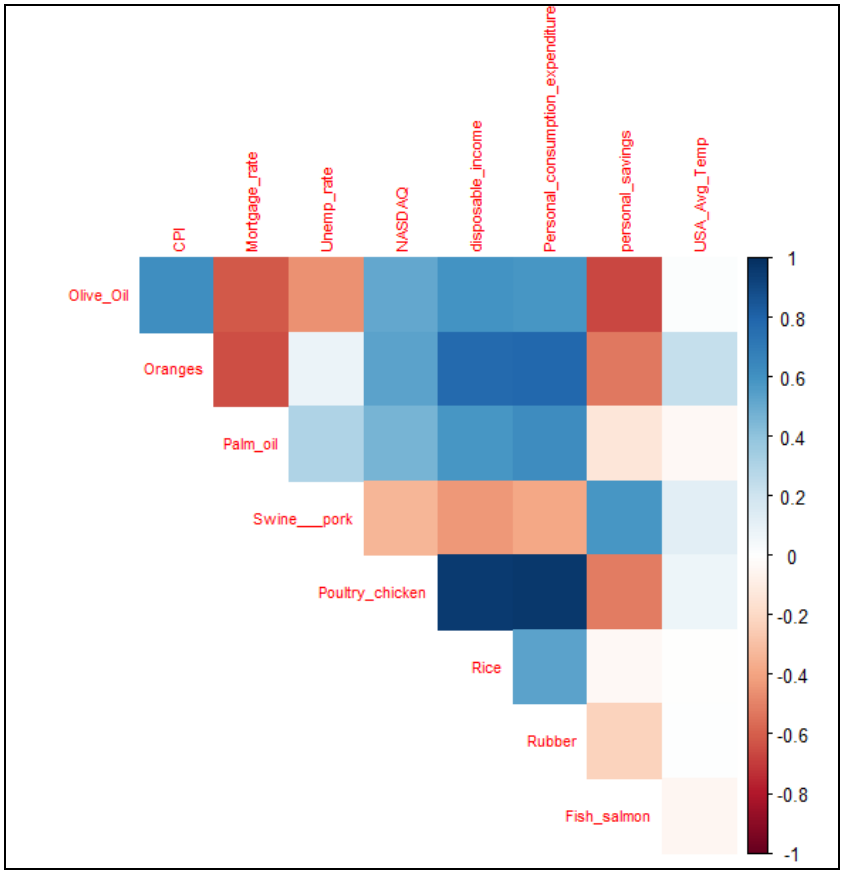
The main activities in pre-processing were data cleansing and merging the three datasets (US commodity prices, US macroeconomic data, and US weather data). We identified that the weather data was not sufficiently presenting the required data for the USA. Hence, we had to deviate from the primary source and find another dataset (see description above). As all three datasets include monthly data and the time overlap is November 1980 to February 2016, we merged the datasets accordingly.

Then, we transformed the data in time-series format, either to classes “ts” or “xts”/”zoo”, to utilise the manipulation benefits of such formats as well as to satisfy the requirements

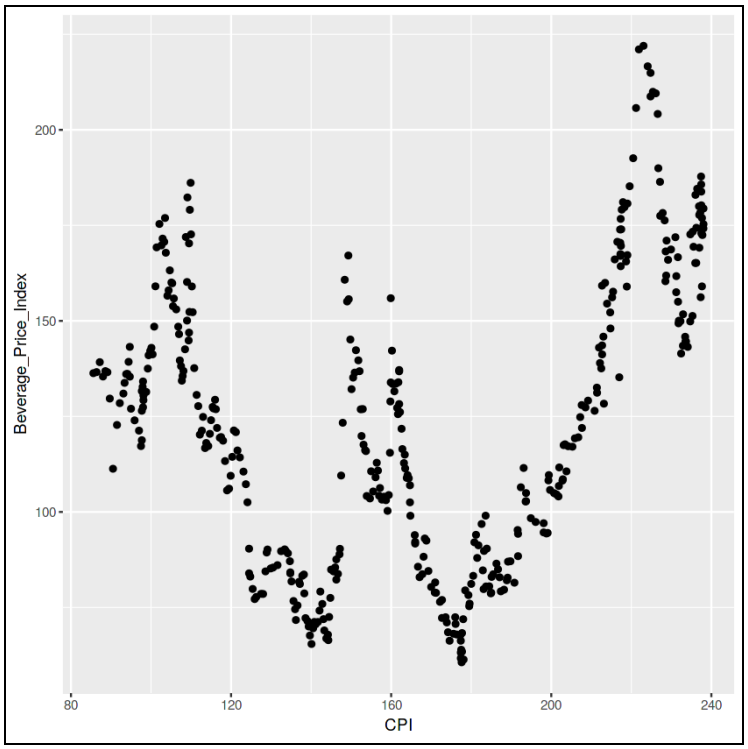
[3] On https://www.kaggle.com/datasets/sarthmirashi07/us-macroeconomic-data
[4] On https://www.researchgate.net/publication/5035085_Modelling_Inflation_Dynamics_A_Critical_Review_of_Recent_Research
[5] On https://medium.com/@ian.colbert711/can-you-predict-commodity-stocks-using-a-weather-forecast-80751dabb36b
[6] On https://onlinelibrary.wiley.com/doi/10.1002/sae2.12041
[7] On https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980020306

of input formats for certain modelling packages. The dataset of avg. temperatures/state was transposed in typical row=datapoint & column = variable format - for easier access.

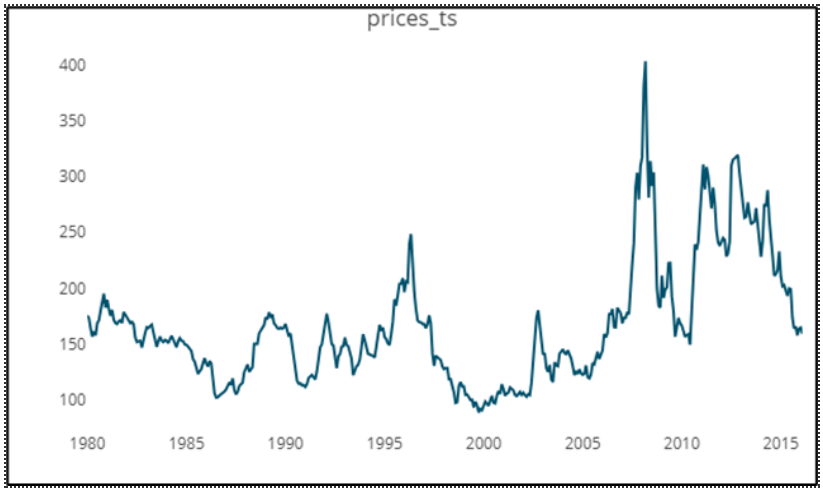
In order to get an understanding of how the dependent variables (commodity prices) and the predictors (macroeconomic and weather data) are correlated, we created a correlation matrix. Here we show one example, were we identified that, for example, pork prices are negatively correlated to personal consumption but poultry prices are highly positively correlated, suggesting that when personal consumption expenditure is increasing, people tend to buy more poultry and less pork.



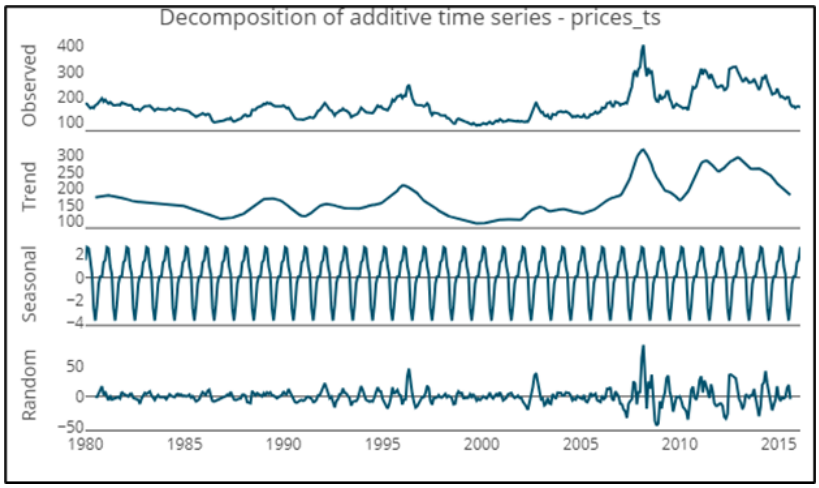
The correlation matrix indicated that the variables "Beverage_Price_Index," "Industrial_Inputs_Price_Index," "Agricultural_Raw_Materials_Index," and "Metals_Price_Index" show the highest correlation. To further investigate the relationship, we specifically examined the dependency of the Beverage price index with the CPI (Consumer Price Index), as shown in the following graph:



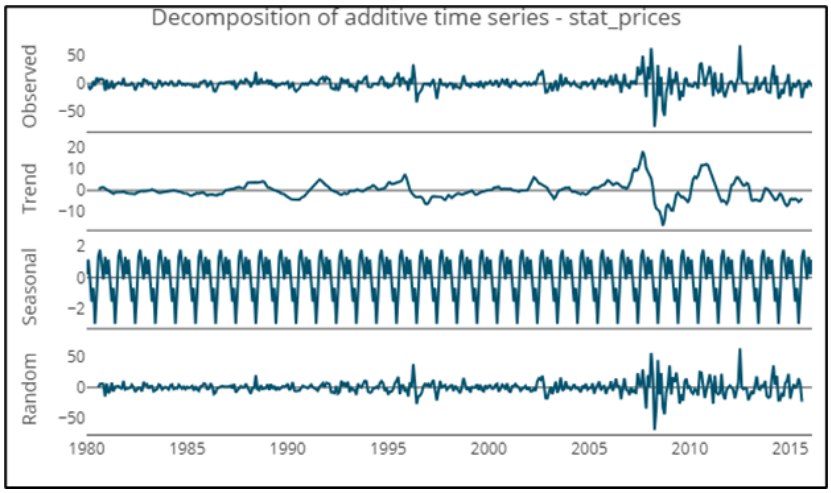
As a next step we went into detail by doing some basic charting, for example, wheat prices, to observe the general trends of the price movement:



This led to the exploration of elements of time-series data, namely trends, noise and seasonality. For that purpose, decomposition plots were utilised, as below, again for wheat price:



It was noted further that the trend component of the wheat price series fluctuates over the various periods, leading to the conclusion that the dataset might be non-stationary. This might be acceptable for certain models, but for others not; thus, the optional transformation was adopted by taking the difference from the series and stabilising it. The Augmented Dickey-Fuller (ADF) test was applied on series before and after treatment to reconfirm the observation from the decomposition graphs, as noted above. The decomposition graph of the series passing the ADF test is charted below:



The same tests were run on all data - both temperature datasets were stationary.

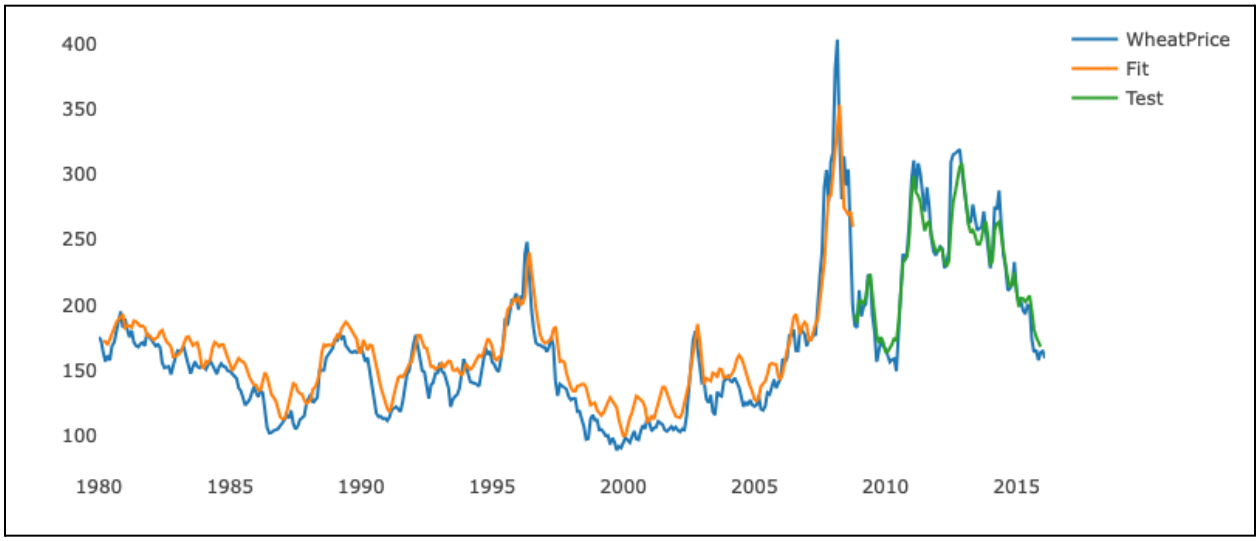
PRIMARY MODELS DEVELOPED

Several models were developed:

- The LSTM (Long Short-Term Memory) model - which leverages Keras and TensorFlow modules via package 'TSLSTM'. Here, the price of wheat was predicted based on temperature values in all states as well as temperature values in wheat-producing states only, i.e. Kansas (producing ca 25% of US wheat). The model yields MAPE = 0.08, with ca. 60 k parameters and reasonable compute time.
- Grangertest were performed to look for likely relations between weather parameters (precipitation and temperatures) and the prices of commodities. This was done for a timeframe from January 1980 to January 2016.
- ARIMA modelling was performed to fit the relations between the weather parameters and the commodity prices. This was also done for a timeframe from January 1980 to January 2016.

Results/Discussion

LSTM model fitted to 80% of data yielded MAPE of 0.08 and RMSE 23.14 on the test set. The series was lagged by 2 and scaled automatically via package. Various setups of parameters were used, varying epochs and number of units in the LSTM layer, as well as utilizing regularisation via dropout rate in the range of 0.5-0.8. Future efforts will be invested in expanding the forecasting horizon as well as introducing other commodities for forecasting. Graph of actual wheat price as well as descaled fit & predicted (test) values of LSTM model:



Granger tests were performed to indicate whether there is a statistically significant relation between the weather effect and the commodity price. This does not indicate

causation. But we did try to find parameters for which it would make sense the relation would be causal. The fuel price index was found to be statistically significantly related to the average monthly temperature in New York. Soybean meal prices were compared to factors in Aledo, Illinois, this town is on the border of the two states with the highest soybean production in the USA. [8] In this case, the average minimum temperature was also found to have a statistically significant correlation with the soybean price.

Commodity predicted for	Weather influence used as predictor	P value	Lag used
Fuel energy price index	Daily minimum temperature in averaged by month in New york	0.045 *	1 month
Fuel energy price index	Daily minimum temperature in averaged by month in New york	0.149	12 months
Soybean meal	Average precipitation in Aledo, Illinois	0.42	3 months
Soy bean meal	Daily minimum temperature in averaged by month Aledo, Illinois	0.016 *	3 months
Soy bean meal	Daily minimum temperature in averaged by month in Aledo, Illinois	0.078	6 months

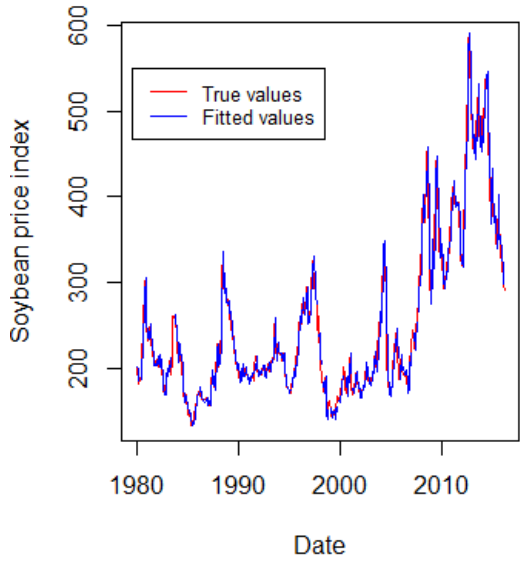
ARIMA models fitted:

Commodity	Weather influence used	ARIMA type	Log likelihood without weather influence	Log likelihood with weather influence	Trending factor
Fuel price index	Average monthly temperature New York	(1,1,0)	-961.5	-958.08	0.15$/°F
Soybean price	Average monthly temperature Aledo	(1,1,0)	-1848.19	-1760.77	0.20$/°F

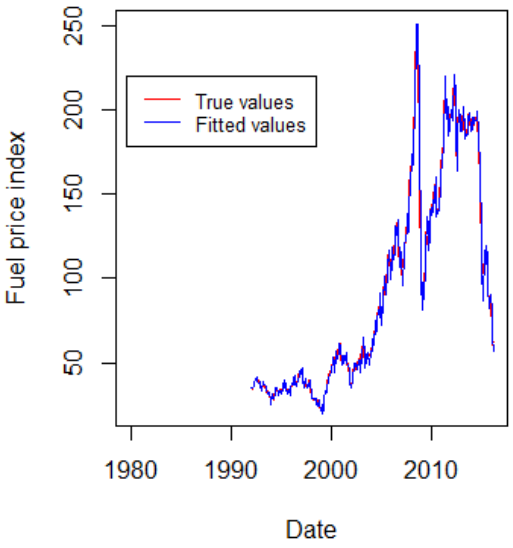
Both models improve the log likelihood by a little amount. The trending factors are also

[8] https://www.cropprophet.com/soybean-production-by-state-top-11/

very small looking at the overall variance. This is illustrated best by the following two figures:



Fitted and real price data for the prediction of the soybean price index using average monthly temperatures



Fitted and real price data for the prediction of the fuel price index using average monthly temperatures

The factor for the ARIMA model for fuel price index is positive, this seems counterintuitive and will need to be investigated further. Lower temperatures would be expected to increase the fuel price due to the higher demand for heating. For the soybean index, this could be explained by positive and negative factors. In this case, it is also positive.

NEXT STEPS

At this moment, many models have been fit, and more are being fit. In the coming phase, we will have to compare the results of the various models. For this, we will have to find a way to do this in a meaningful way by comparing the same success factors.

REFERENCES

See footers