PREDICTING COMMODITY PRICES

Predicting Agricultural Commodity Prices between November 1980 and February 2016 using a Deep Learning Approach with Macroeconomic and Climate Data between November 1980 and February 2016

Team 9

July 20th, 2023

MGT6203, Summer 2023

Rade Bajic, Eddie Morrissey, Stijn Jorissen, Nurly Kuzdikbay, David Basler

1. INTRODUCTION

This study found that state of the art machine learning algorithms are not able to forecast agricultural commodity prices reliably, but can provide insights in what directions prices will go in the future. This is already important from a business perspective, as commodity traders for example are highly interested in predictive models that process information and return an indication of where future prices are heading. We applied algorithms such as ARIMA, LSTM, RNN, Silverkite, and others to come to our conclusion.

We worked with monthly time series data from November 1980 to February 2016. By examining data preprocessing, model selection, training, and evaluation, this report aims to shed light on the efficacy and applicability of ML models in this particular domain. The findings and insights derived from this research can contribute to enhancing decision-making processes and strengthen the understanding of how commodity prices relate to different exogenous variables.

2. LITERATURE REVIEW

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 Furthermore, most books on the topic were published +5 years prior, which makes them outdated in both the coding approaches applied as well as models proposed.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)[1]
- Can you predict commodity stocks using a weather forecast? by I. Colbert Medium[2]
- Climate and environmental data contribute to the prediction of grain commodity prices using deep learning[3]

[1] On https://www.researchgate.net/publication/5035085_Modelling_Inflation_Dynamics_A_Critical_Review_of_Recent_Research

[2] On https://medium.com/@ian.colbert711/can-you-predict-commodity-stocks-using-a-weather-forecast-80751dabb36b

[3] On https://onlinelibrary.wiley.com/doi/10.1002/sae2.12041

- Price forecasting and evaluation: An application in agriculture. By Jon A. Brandt, David A. Bessler[4]

3. METHODOLOGY

3.1. HYPOTHESIS

The main hypothesis is: There is a way to predict commodity prices better than ARIMA using state of the art machine learning algorithms (such as Recurrent Neural Networks, Long Short-Term Memory Neural Networks).

The most important sub-hypothesis is: Macroeconomic data and weather will be a factor in the prediction of the prices of commodities.

Following the above, another sub-hypothesis is: that Transformer models can outperform direct models such as LSTM and traditional models such as ARIMA.

3.2. DATA DESCRIPTION

“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 utilized soybean, corn, wheat and oil prices in various models in an attempt to predict the price movements[5].

“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 centers of each state.[6]

“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 1980 to May 2022. The data is from Federal Reserve Economic Data and retrieved from

[4] On https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980020306

[5] On https://www.kaggle.com/datasets/vagifa/usa-commodity-prices

[6] On https://www.kaggle.com/datasets/justinrwong/average-monthly-temperature-by-us-state

Kaggle.[7]

Data from the national weather service was used to acquire precipitation data from the US. Aledo, IL was used for Illinois [8]and Wichita, KS was used for Kansas [9].

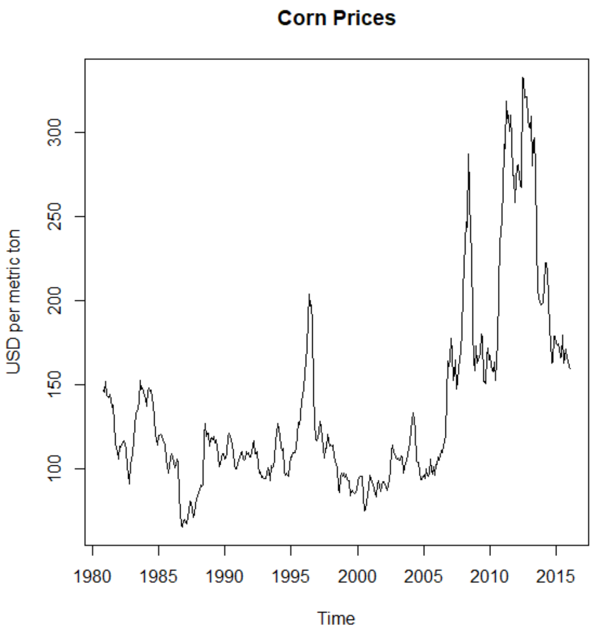
3.3. DATA CLEANING AND PREPROCESSING

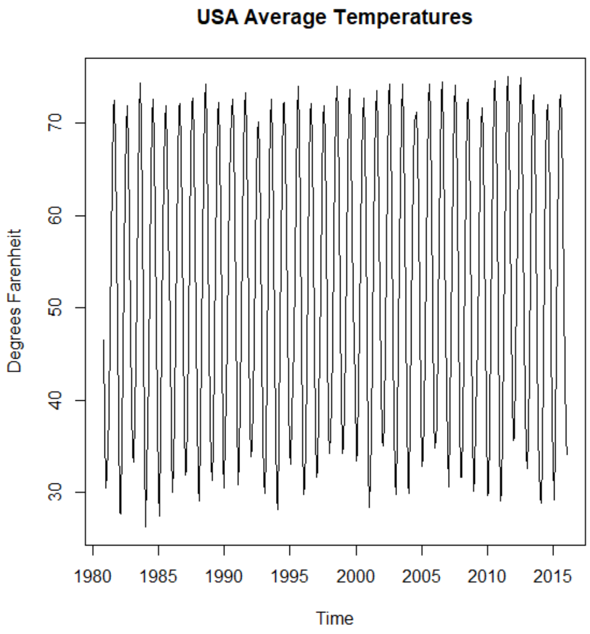
The main activities in pre-processing were data cleansing and merging the three datasets (US commodity prices, US macroeconomic data, and US weather data). We identified that the weather data was not sufficiently presenting the required data for the USA. Hence, we had to deviate from the primary source 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 utilize the manipulation benefits of such formats as well as to satisfy the requirements of input formats for certain modeling packages. The dataset of avg. temperatures/state was transposed in typical row=datapoint & column = variable format - for easier access.

Inspection of corn prices
The USA is the largest corn producer in the world in terms of total production, the majority of which is grown in the Heartland region. Iowa and Illinois, the top corn-producing States, typically account for about one-third of the US crop.
We analyzed the movements of the corn price over the observation period and identified no specific pattern regarding seasonality or linear trends. The lack of seasonality can best be shown in comparison to the average temperatures where the effects of the four seasons are obvious (see figures below). Regarding the trend characteristic we identified an increase of corn prices with a significant change of variability over time, implying a non-stationary time series. These characteristics were expected, as the trade of commodities is embedded in a complex system where prices depend on many factors. Looking at the development over time, we recognized that around the financial crisis in 2008, the volatility of the price increased significantly.

[7] On https://www.kaggle.com/datasets/sarthmirashi07/us-macroeconomic-data
[8] https://www.ncei.noaa.gov/cdo-web/datasets/GSOM/stations/GHCND:USC00110072/detail
[9] https://www.ncei.noaa.gov/cdo-web/datasets/GSOM/stations/GHCND:USW00003928/detail





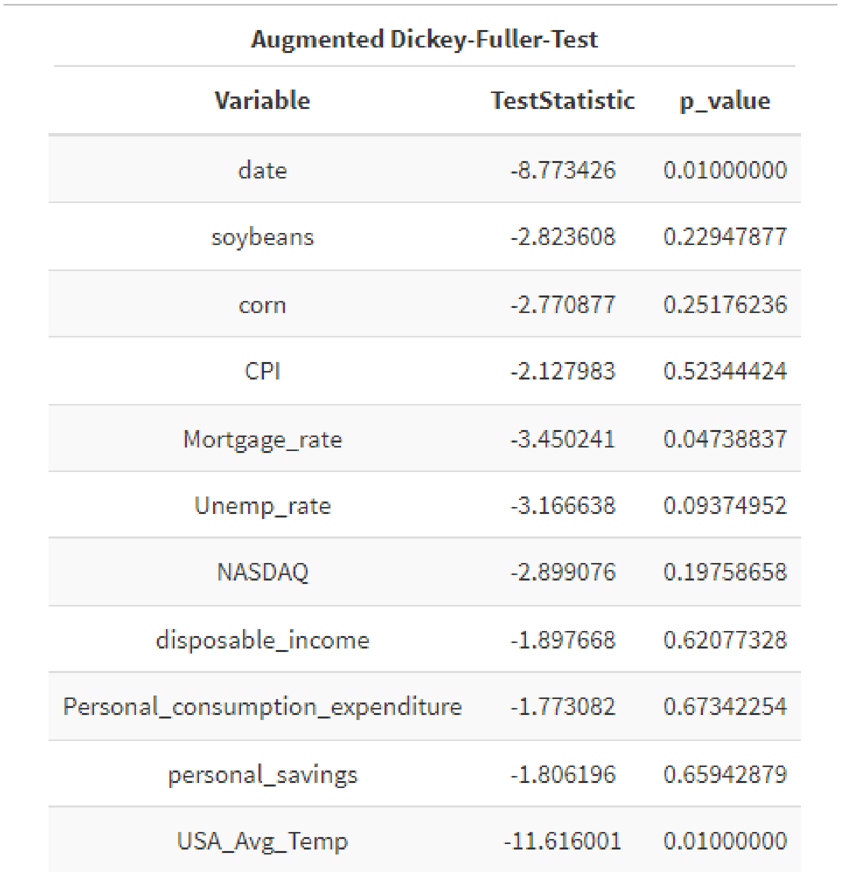
Test for stationarity

We observed in the figure above that the corn prices might be non-stationary in itself. We tested the dependent and independent variables for stationarity, to ensure all variables are stationary or appropriately transformed to achieve stationarity before proceeding with modeling. By doing so, we reduced the risk of including variables that may introduce bias or produce unreliable results in the final model.

An appropriate test is the Augmented Dickey-Fuller-Test (ADF) which is a statistical test with a null hypothesis that a unit root is present in the time series sample (non-stationary time series), and an alternate hypothesis that no unit root is present in the time series sample (stationary time series). If the null hypothesis fails to be rejected, this test may provide evidence that the series is non-stationary. We tested if the test statistic is smaller than the p-value of 0.05, which would mean the data does not have a time-dependent structure, and is stationary. One important parameter of the ADF test is choosing the lag length "k". We decided for the default value according to the R documentation of the ADF test which corresponds to the suggested upper bound based on a common rule of thumb known as the truncation lag selection criterion. We deem this approach of taking the upper bound was appropriate as it was proven that it is better to error on the side of including too many lags, according to a paper of the University of Washington[10].
Based on the table below, we concluded that for most of the variables we cannot reject

[10] On https://faculty.washington.edu/ezivot/econ584/notes/unitrootLecture2.pdf

the null hypothesis, and the data is mainly non-stationary. However, for the variables Mortgage Rate and USA Average Temperatures we reject the null hypothesis (p-value below 0.05) and conclude that the data is stationary.



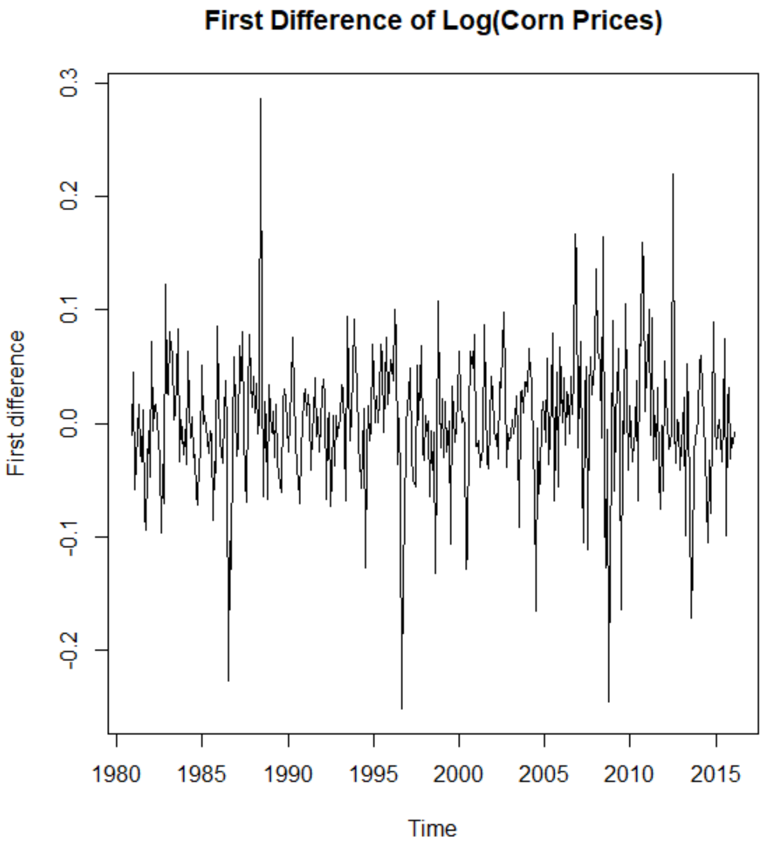
Augmented Dickey-Fuller-Test

Variable	TestStatistic	p_value
date	-8.773426	0.01000000
soybeans	-2.823608	0.22947877
corn	-2.770877	0.25176236
CPI	-2.127983	0.52344424
Mortgage_rate	-3.450241	0.04738837
Unemp_rate	-3.166638	0.09374952
NASDAQ	-2.899076	0.19758658
disposable_income	-1.897668	0.62077328
Personal_consumption_expenditure	-1.773082	0.67342254
personal_savings	-1.806196	0.65942879
USA_Avg_Temp	-11.616001	0.01000000

Transform the time series

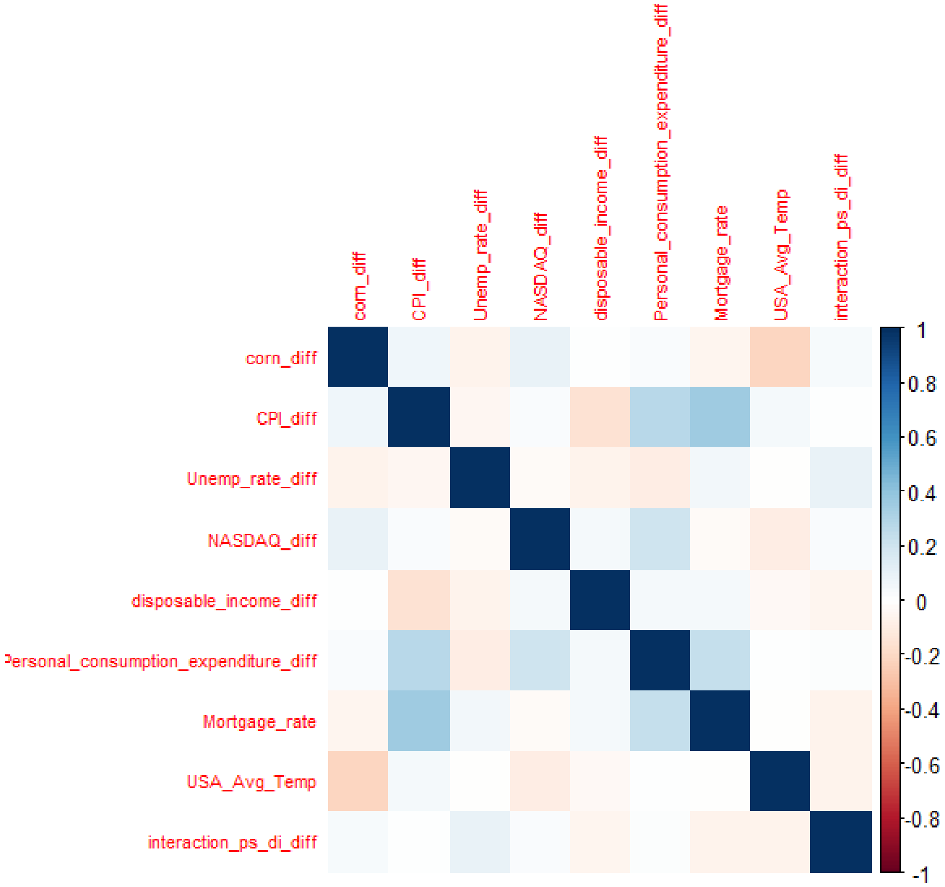
Once non-stationarity for the variables mentioned was defined, we applied transformations to make the time series stationary. Hence, we created the first differences of the logarithmized series for the respective variables. One important step was to make sure there are any negative or NULL values that would distort the process.

As an example, see the transformed corn prices below. The figure shows the similar pattern as the corn price figure above, where we already identified increased variability around the financial crisis in 2008. Due to this transformation the trends and seasonality were removed and the variance of the series was stabilized.



Test for causality

The transformed variables were merged to the stationary data and tested for granger causality._This test is used to assess the causal relationship between time series. It assumes that the independent variables in the model are not linear combinations of each other. We identified some perfect correlations within the macroeconomic variables between for example personal savings and disposable income. This makes sense, as you would intuitively assume that if someone has more disposable income the personal savings increase. Thus, we created an interaction term by multiplying the two highly correlated variables together. An interaction term can help capture the joint effect of the original variables. Subsequently, we dropped the original variable of personal savings in order to eliminate the multicollinearity.



On the basis of the variables shown above and any other relevant iteration, we performed the Granger causality test. We tested different lags and looked for variables that have a significant p-value (<0.05). Correlations that were investigated are showcased in the following table below, the statistically significant correlations are highlighted with a star.

Predictor	Predicted	Lowest found p value	Lag for lowest found p value (months)
Temperature in Kansas	US wheat prices	0.21	3
Precipitation in Kansas	US wheat prices	0.57	1
Temperature in New York	US LNG prices	0.10	13
Temperature in New York	US energy price index	0.01*	1
Temperature in Illinois	US soybean price	0.006*	6
Precipitation in Illinois	US soybean price	0.41	3

CPI difference	Corn price	0.036*	1
Unemployment rate	Corn price	0.665	1
Nasdaq difference	Corn price	0.331	1
Disposable income difference	Corn price	0.174	1
Average US temperature	Corn price	0*	1
Mortgage rate	Corn price	0.045*	6
Interaction PS difference	Corn price	0.033*	6

Variation in P values of some of the tests dependent on the lag are plotted in appendix 1.

3.4. DATA MODELLING

The chapter "Data Modeling" covers key topics essential to effective data analysis. It explores feature selection, data splitting, model architecture, model evaluation, hyperparameter tuning, prediction, and model iteration. This chapter equips readers with the knowledge and tools needed to re perform our results.

Feature selection

Based on the results of the Granger test, we performed feature selection of significant variables with p-value smaller than 0.05. Furthermore, we limited the selection to the three most important features. The major analysis was performed using Temp (USA average temperatures), CPI (consumer price index) and Ind_PS_DI (indicator variable personal savings multiplied by disposable income).

Data splitting

One common approach to split data is to use a holdout validation strategy where you set aside a specific number of the most recent months as the test set, and use the remaining data for training. This ensures that the model is trained on historical data and evaluated on more recent observations. We decided to use the most recent 48 months as a test set.

TFT particular features

Tests were done and error in the TFT package was detected - it relies on “tsibble” to understand frequency of the time-series, however, freq = month was not correctly detected by tsibble, causing issues with recipe formulation and forecasting. Thus, time-series were “upsampled” from month to week, by recycling of index and imputation of monthly values to adjoining weeks. Furthermore, the TFT model requires 3 types of variables - “index”, “key” and “unknown”, while it allows for optional “known” variable. Others[11] have shown that introduction of categorical predictor (“known”) variables can leverage otherwise undetected / un-understood long-term dependencies. In this case, we’ve selected “known” = day, day of the week, month, and ordinal number of days since beginning of time-series.

Data architecture

Building the LSTM and RNN model required tuning of many hyperparameters such as the amount of lags, epochs. We tested different variations and kept the best performing model. Notably, the LSTM package in Python was not able to consider exogenous variables and could only perform a prediction on the chosen time series itself. However the remaining algorithms were able to base their prediction on the provided independent variables.

Model evaluation

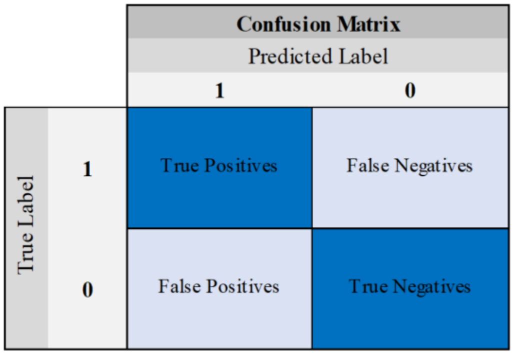
We decided to use the following metrics to compare and assess our models:

1. MAPE (Mean Absolute Percentage Error)
    a. MAPE measures the average percentage difference between the actual values and the predicted values. It is calculated as the mean of the absolute percentage errors over all test set samples.
2. RMSE (Root Mean Squared Error)
    a. RMSE measures the root mean squared difference between the actual values and the predicted values. It is calculated by taking the square root of the mean of the squared errors.
3. R2 (R-squared)
    a. R-squared, also known as the coefficient of determination, measures the proportion of the variance in the target variable that is predictable from the predictor variables. It ranges from -∞ to 1, and negative values can occur when the model performs worse than a horizontal line. A value of 1 indicates a perfect fit.

In addition, we performed a classification of the predictions. Classification problems are usually summarized in a four-celled matrix (see below). The matrix compares actual vs. predicted results and is the basis to calculate meaningful ratios. Dark-blue shading refers to the number of correct prediction counts, and light-blue shading refers to incorrect prediction classifications. Adding the raw classifications of each cell gives the total number of observations.

[11]

- True Positives (TP): Number of correctly predicted targets (i.e., correctly predicted delinquent borrowers)
- True Negatives (TN): Number of correctly predicted non-targets (i.e., correctly predicted non-delinquent borrowers)
- False Positives (FP): Number of falsely predicted targets (i.e., falsely predicted delinquent borrowers)
- False Negatives (FN): Number of falsely predicted non-targets (i.e., falsely predicted non-delinquent borrowers)



		Confusion Matrix	
		Predicted Label	
		1	0
True Label	1	True Positives	False Negatives
	0	False Positives	True Negatives

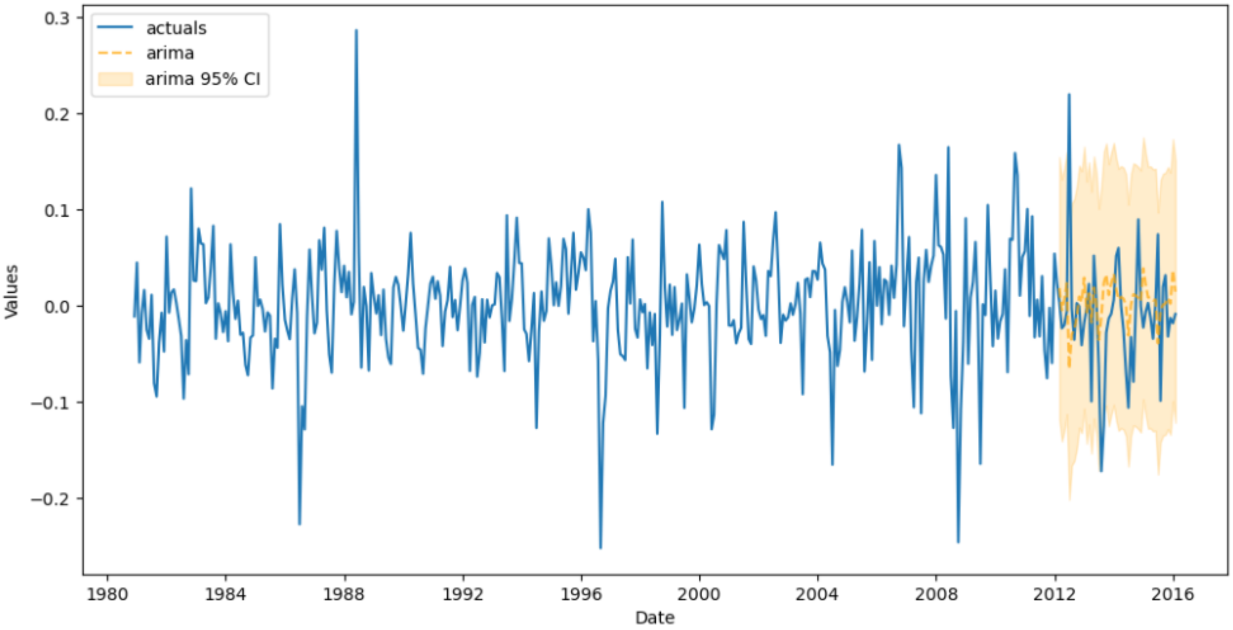
From this basis, additional metrics such as accuracy, precision, recall, or F1 score can be derived. However accuracy is the best metric for this particular use case, because you want to predict both the increase or decrease of price with the same probability.

4. EMPIRICAL ANALYSIS

Corn price prediction

Due to stationarity the tested models tried to predict the logarithmic difference of corn prices (hereafter named “Corn Prices”). The results are sobering as the results for all four models were unsatisfactory. However, when thinking intuitively about commodity prices, it is reasonable that three variables Temp, CPI, Ind_PS_DI (see description above) are not enough to predict prices in a complex system which includes speculation, geopolitics, and other factors.

The best prediction of Corn Prices on the test set was observed by the ARIMA model (see below). The confidence interval of 95% captures all observations of the test set except two extreme variations.

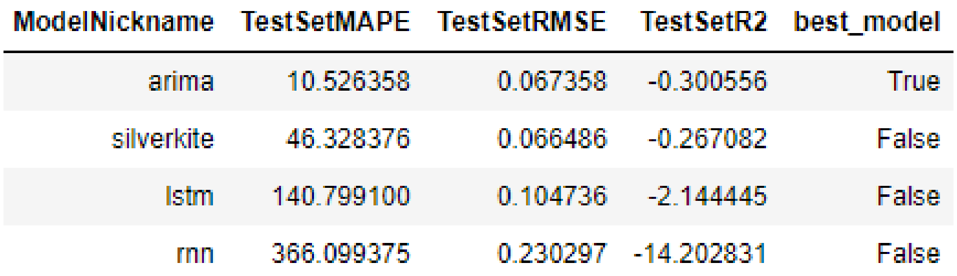


A closer look at the actual predictions (dotted line) reveals that the forecast is not very accurate and ARIMA expects much smoother movements.

Comparing the four models we created a table for comparison (see blow). The ARIMA model's predictions have an absolute percentage error of 10.53% (MAPE). Lower MAPE values indicate better predictive accuracy, and a value close to 0% would indicate a perfect match between prediction and actual values. We see that in this case the ARIMA model was by far the best predictor.

Regarding RMSE, the ARIMA model's predictions deviate from the actual values by 0.0674 units. Lower RMSE values indicate better predictive accuracy, and a value of 0 would again indicate a perfect match between predictions and actual values. In this metric the Silverkite model was slightly better than ARIMA.

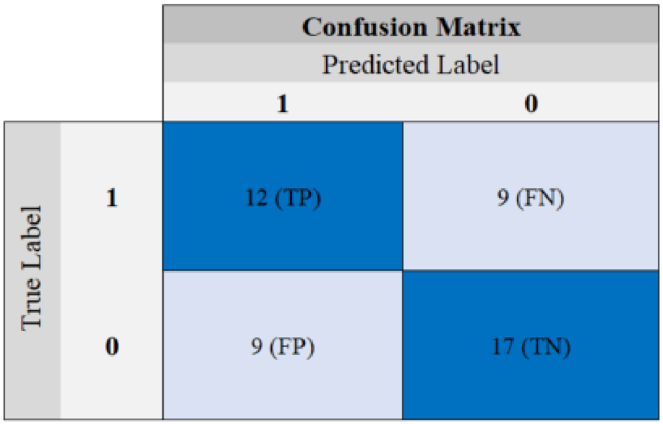
Finally, looking at R-squared, we see that the ARIMA model is performing worse than a horizontal line (a model that predicts the average value of the target variable). This indicates that the ARIMA model is not performing well in capturing the variance in the data and is not a good fit for the given dataset.



ModelNickname	TestSetMAPE	TestSetRMSE	TestSetR2	best_model
arima	10.526358	0.067358	-0.300556	True
silverkite	46.328376	0.066486	-0.267082	False
lstm	140.799100	0.104736	-2.144445	False
rnn	366.099375	0.230297	-14.202831	False

Now, thinking about business justification, a commodity trader indeed dreams of the perfect price predictor engine but she would also welcome it if a model is able to predict the price direction of the next month. Consequently, we deepened the evaluation of our models and checked if prices for the next month are predicted to stay stable or increase (>=0), or if they are predicted to decrease (<0). The classes were named 1,0 (increase, decrease) and compared to the actual movement of the price.

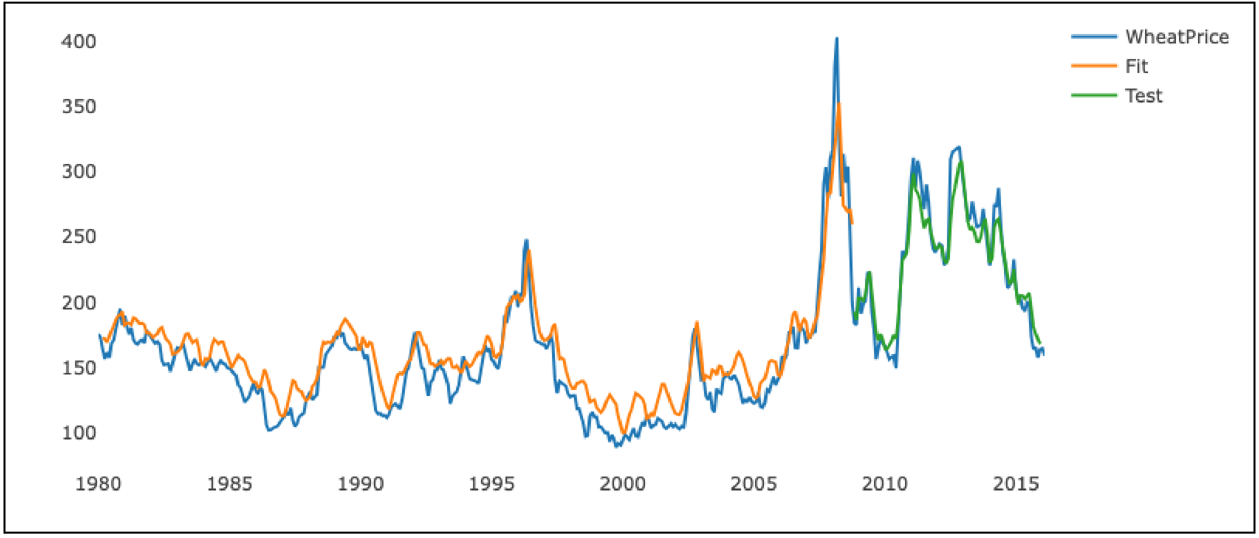
We found that the Silverkite model returned the highest accuracy with 62%. Accuracy is defined by the sum of all correctly predicted directions divided by all predicted directions. Consequently, the Silverkite model is able to predict better than a coin flip. This is a new realization, as the R2 metric in the table above suggested that a horizontal line would better be a better predictor.



		Confusion Matrix	
		Predicted Label	
		1	0
True Label	1	12 (TP)	9 (FN)
	0	9 (FP)	17 (TN)

Prediction of wheat prices based on temperature in Kansas

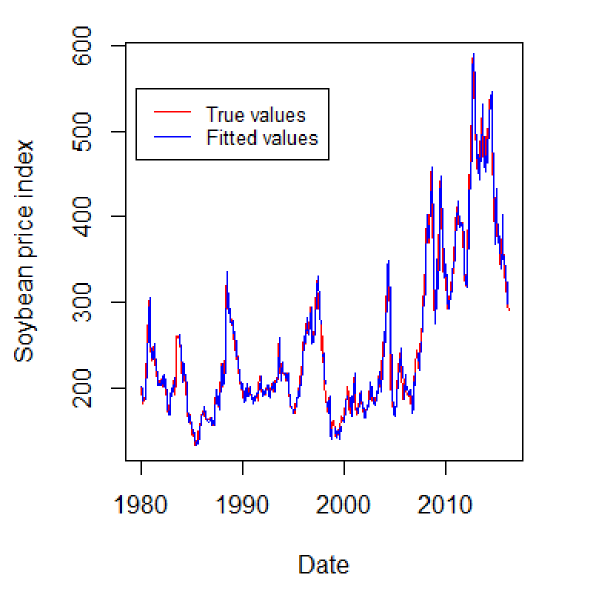
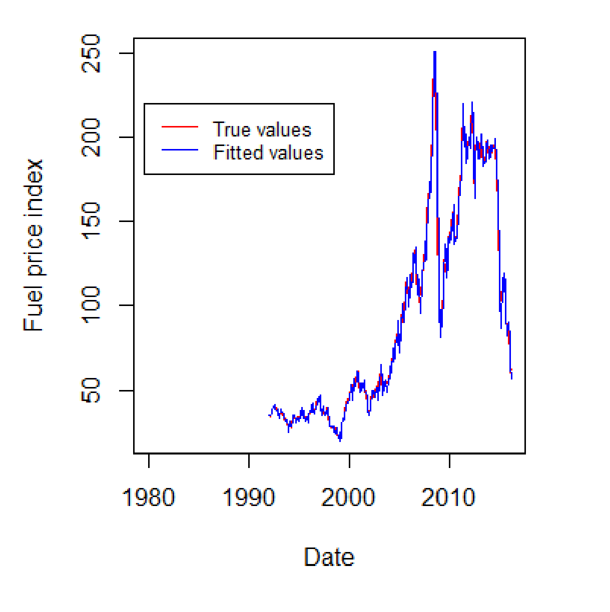
The LSTM model fitted to 80% of data yielded MAPE of 0.08 and RMSE 23.14 on the test set. The series was lagged by 2 and scaled automatically via package. Various setups of parameters were used, varying epochs and number of units in the LSTM layer, as well as utilizing regularization via dropout rate 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:



Nevertheless, the results could not be recreated in Python and thus remains to be further explored what might be the source of such discrepancy i.e. implementation or else.

Further ARIMA models

The relations between the US fuel price index and temperature in New York and between the US soybean price and the average temperature in Illinois were found to be statistically significant. Thus ARIMA models were made to investigate this further.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>Fitted and real price data for the</td>
<td>Fitted and real price data for the</td>
</tr>
</table>

prediction of the soybean price index using average monthly temperatures in Illinois	prediction of the fuel price index using average monthly temperatures in New York

The results for this are shown in the plots above, the parameters can be seen in the prediction summary table. The pdq values for the ARIMA models were determined by choosing the ones which lead to the best fit based on mean absolute error and mean error. The factor for the ARIMA model for fuel price index is positive, this seems counterintuitive and thus the correlation found in the granger test is more likely to be a coincidence than actual causation. 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, thus the prices of soybeans seem to increase with increasing average monthly average temperature in Illinois.

Temporal Fusion Transformers model

An additional model was developed, using Temporal Fusion Transformers[12] architecture. This approach is in line with current "state-of-the-art" neural networks for forecasting of time-series i.e. utilization of attention mechanisms for inference of long-term dependencies. We have decided to set number of epochs to range 1 - 20, in line with observations[13] that large datasets and loss function optimization can safely reduce compute time/costs. We have utilized the R-package "mlverse/TFT"[14] which offers R implementation of Temporal Fusion Transformers and is constructed on the basis of the referred paper (see footnotes 12 & 14). In line with suggestions that TFT outperforms other forecasting models like: *"iterative methods (e.g., DeepAR, DeepSSM, ConvTrans) and direct methods (e.g., LSTM Seq2Seq, MQRNN), as well as traditional models such as ARIMA, ETS, and TRMF"*[15] - we also test this as a sub-hypothesis of this report. Several test runs were made. The package offers a wide range of tuning methods, for example "learning rate" optimization via graphing and auto-validation. However, the training process is extremely compute-heavy and thus a limited number of epochs were utilized. Thus, the model did show better performance than others but was not the best.

[12] https://doi.org/10.48550/arXiv.1912.09363

[13] https://doi.org/10.48550/arXiv.1906.06669

[14] https://mlverse.github.io/tft/index.html

[15] https://ai.googleblog.com/2021/12/interpretable-deep-learning-for-time.html#:~:text=any%20time%20step.-,Forecasting%20Performance.-We%20compare%20TFT

Summary of models fitted for prediction of absolute values:

Commodity	**Weather influence used**	**Model type**	**MAPE**	**RMSE**	**Trending factor**
Corn price	US average temperature	Arima(2,0,1)	10.53	0.67	-
Corn price	US average temperature	Silverkite	46.33	0.066	-
Corn price	US average temperature	LSTM	140.8	0.1	-
Corn price	US average temperature	RNN	366.1	0.23	-
Fuel price index	Average monthly temperature New York	ARIMA(0,1,0)	12	1.09	0.0006 $/°F
Soybean price	Average monthly temperature Illinois	ARIMA(0,1,0)	10	2.61	0.0019 $/°F/ton
Wheat price	Average monthly temperature Kansas	LSTM	0.08	23.14	-
Wheat price	Average monthly temperature Kansas	TFT	1.96*	1.23	

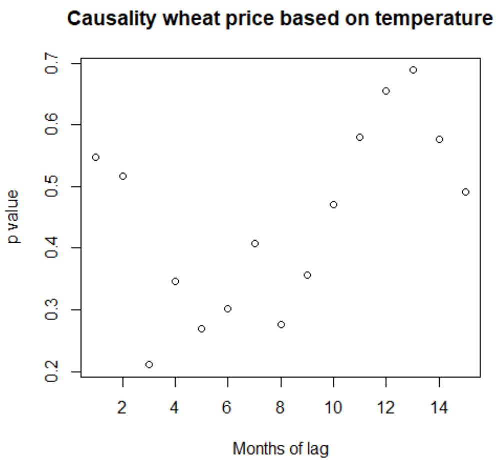
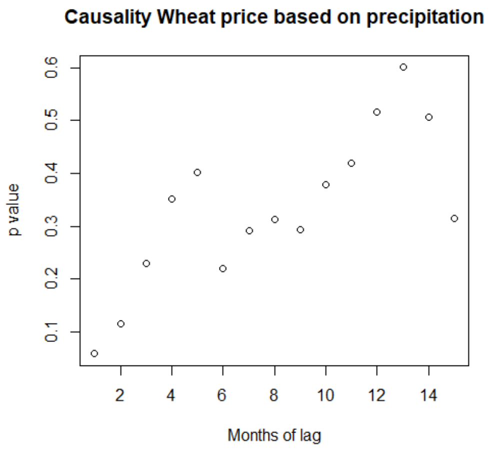
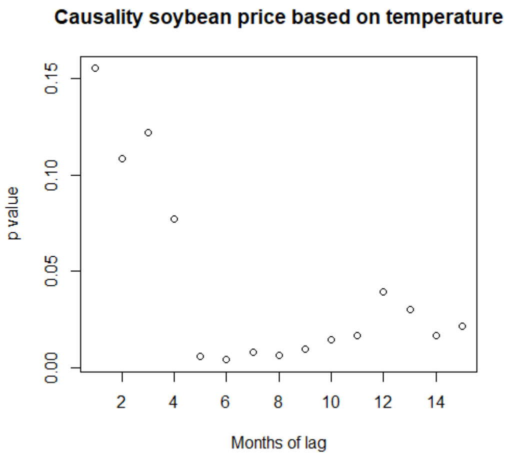
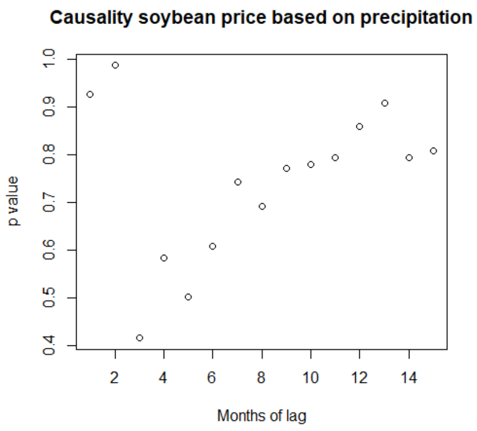
CONCLUSION

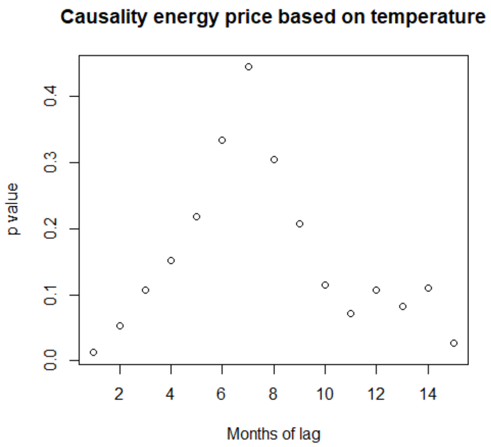
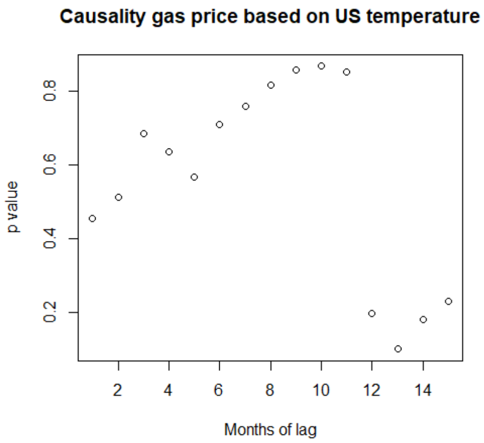
Various models were made to predict commodity prices based on weather influences, primarily monthly temperatures. Arima models lead to extremely small decreases in the mean average error and the mean error. Also the effect size of the external regressor was extremely small. Thus even though statistically significant relations were found (p<0.05) it is questionable if this is due to chance or if the effect is really there for the Arima models. It is especially unlikely that causality was found in these cases. Transformer based model did not show stelar performance, however, it should be noted that there was limitation in available compute resources.

From our study we suggest that further research should focus on predicting the price direction of commodity prices rather than the price itself. This statement is based on the findings that commodity prices reveal a lot of random variability that cannot be explained in our study. Hence, a model is maybe not able to predict the exact price, but giving a reliable direction of where prices might go is immensely valuable for the trading system. We would recommend extending the exogenous variables considered in our study using for example macroeconomic data from other countries that are large producers of agricultural commodities (e.g. Ukraine for wheat, China for rice and corn).

Appendix 1: casualty test plots

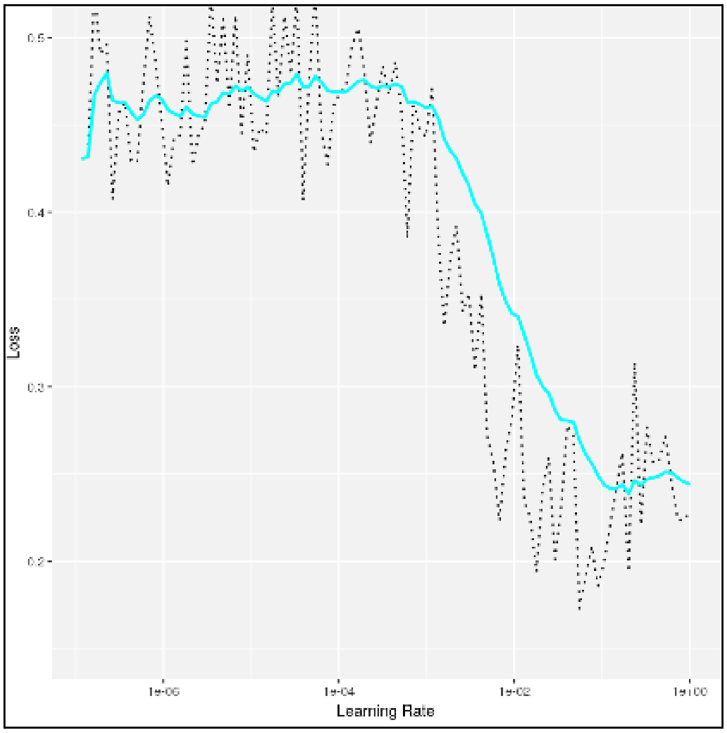
In this section the causality test plots are shown. The causality tests were done over a lag period of various months to see if the causality would be higher at a specific lag.

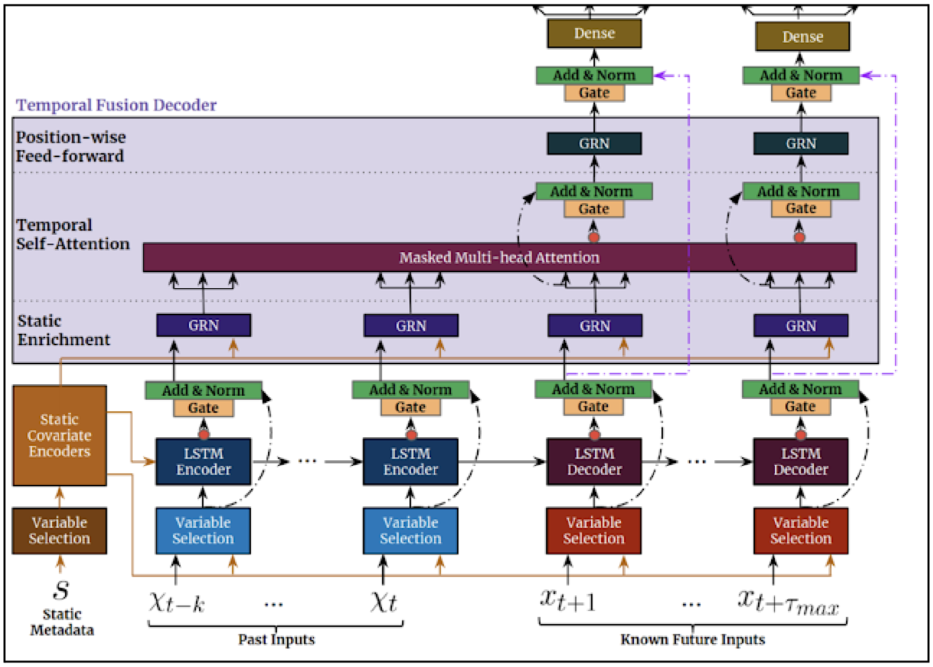
 Causality wheat price based on temperature 	 Causality Wheat price based on precipitation
Causality test between temperature in Kansas and the wheat prices in the US. No statistically significant correlation was found with the Granger test.	Causality test between precipitation in Kansas and the wheat prices in the US. No statistically significant correlation was found with the Granger test.
 Causality soybean price based on temperature 	 Causality soybean price based on precipitation
Causality test between temperature in Illinois and the soybean prices in the US. Various statistically	Causality test between precipitation in Illinois and the soybean prices in the US. No statistically

significant relations were found.	significant correlation was found with the Granger test.
	
Causality test between temperature in New York and the energy prices in the US. A statistically significant relationship was found with a monthly lag of 1 month.	Causality test between temperature in New York and the gas prices in New Orleans terminals. No statistically significant correlation was found with the Granger test.

Appendix no. 2 TFT plots & stats (loss, errors, model structure)

In this section, adjoining plots of the TFT model are enclosed.





https://ai.googleblog.com/2021/12/interpretable-deep-learning-for-time.html