Using ML algorithms to Predict Commodity Price Trends

Cai Jianlin 3035500970
Wong Sai Kit 3035475307

Jan 2022

Intermediate Report

Group fyp21074
Supervisor: Dr Schnieders Dirk
Submission Date : Jan 2022
## Contents

1 Abbreviation .................................................. 1

2 Acknowledgment ............................................... 2

3 Abstract .......................................................... 2
   3.1 Definition of Commodities .......................... 2
   3.2 Selection of Commodities ......................... 2
   3.3 Related Studies ......................................... 3
   3.4 Calculation of commodities Prices ............... 3
   3.5 Dynamic Changes in Pricing ....................... 3

4 Introduction ................................................... 5
   4.1 Background .............................................. 5
   4.2 Importance of Commodities in Daily Life .......... 5
   4.3 Core of the Study ..................................... 5
   4.4 Theories Behind the Data Analytic ............... 6

5 Prior Work ...................................................... 7
   5.1 How the Results of Previous Works ............... 7
   5.2 What are the Improvements ....................... 7

6 Methodology .................................................... 8
   6.1 Overview ............................................... 8
   6.2 Web Scraping ........................................... 8
   6.3 How to use Machine Learning Algorithms to predict the results ........................................ 10
   6.4 What are the affecting factors .................... 12
      6.4.1 Climate and Seasonality ....................... 12
      6.4.2 Geopolitics ...................................... 13
      6.4.3 Summary .......................................... 13
   6.5 Relationships of different commodities .......... 13
   6.6 Time Frame ............................................. 14
   6.7 ML models to be tested ............................ 14
   6.8 Neural Network ....................................... 15
   6.9 Web App Structure .................................... 16
   6.10 Final Visualization Version of Predictions ........ 16
# Results

## 7.1 Preliminary ML Model - RNN

### 7.1.1 Software Specifications

### 7.1.2 Data Preprocessing

### 7.1.3 LSTM Layers

### 7.1.4 Prediction Result

## 7.2 Other Models in Testing

## 7.3 Difficulties

## 7.4 Limitation

## 7.5 Conclusion

## 7.6 Future Work

## 7.7 Optional Pathways

### Reference

---

**Table of Contents**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Results</td>
<td>18</td>
</tr>
<tr>
<td>7.1 Preliminary ML Model - RNN</td>
<td>18</td>
</tr>
<tr>
<td>7.1.1 Software Specifications</td>
<td>18</td>
</tr>
<tr>
<td>7.1.2 Data Preprocessing</td>
<td>18</td>
</tr>
<tr>
<td>7.1.3 LSTM Layers</td>
<td>20</td>
</tr>
<tr>
<td>7.1.4 Prediction Result</td>
<td>20</td>
</tr>
<tr>
<td>7.2 Other Models in Testing</td>
<td>22</td>
</tr>
<tr>
<td>7.3 Difficulties</td>
<td>24</td>
</tr>
<tr>
<td>7.4 Limitation</td>
<td>24</td>
</tr>
<tr>
<td>7.5 Conclusion</td>
<td>25</td>
</tr>
<tr>
<td>7.6 Future Work</td>
<td>25</td>
</tr>
<tr>
<td>7.7 Optional Pathways</td>
<td>25</td>
</tr>
</tbody>
</table>

---

8 Reference  27
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A simple web scraping code</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>A simple web scraping code</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Web Scraping Extraction</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Convert Dataframes</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>Merge Frames</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>Correlation Matrix</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>Seasonality</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Neural Network Display</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>Final Prediction of Commodities</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>Splitted gold dataset</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td>Trimmed pandas data input</td>
<td>19</td>
</tr>
<tr>
<td>12</td>
<td>Code for a [0-1] feature transformation</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>Transformation of dataset with a batch size of 6</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>A LSTM working principle graph. Signals are forgotten at the forget gate</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>Gold price prediction result from Dec 2020 to Nov 2021</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>The equations for MAE, MSE and R2</td>
<td>21</td>
</tr>
<tr>
<td>17</td>
<td>An example of including political stability index in the input for RFR model</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>RFR model’s prediction result</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>RFR model’s accuracy test</td>
<td>23</td>
</tr>
<tr>
<td>20</td>
<td>SVR model’s implementation</td>
<td>23</td>
</tr>
<tr>
<td>21</td>
<td>SVR model’s prediction graph</td>
<td>23</td>
</tr>
</tbody>
</table>
1 Abbreviation

CPU  Central Processing Unit

IMF  International Monetary Fund

ML   Machine Learning

ODE  Ordinary Differential Equation

PDE  Partial Differential Equation

RAM  Random Access Memory

USD  US dollar(s)

LSTM Long Short Term Memory

RNN  Recurrent Neural Network

MSE  Mean Square Error

MAE  Mean Absolute Error

RFR  Random Forest Regressor

SVR  Support Vector Machines

GNN  Graph Neural Networks

GANs Generative Advanced Neural Networks
2 Acknowledgment

We are especially thankful to our supervisor Dr. Schnieders Dirk for giving us enormous support and guidance throughout this project.

Also, we would like to express our sincerest gratitude to our English teacher Dr. Locky Law and course coordinator Dr. Ken Ho for providing us with a lot of useful materials and guidelines to follow on to write this report.

3 Abstract

3.1 Definition of Commodities

Commodities are products that can be bought and sold and related to the mainstream of the productions. [17] The 4 main categories of commodities are metals, energy sources, raw foods and plastics. [4] In the study, since it is impossible to cover everything on the market, we narrowed down into 3 categories and 6 commodities based on the “most traded” criteria. It is a reasonable amount of samples to be tested on a local machine and with it we can at the same time consider the correlations between within the same category or between different categories of commodities. [22] For example, the study of correlation of coal and natural gas as they have substitution effect, also coal and metals as they will change in equilibrium prices of each other. [10]

3.2 Selection of Commodities

In the study, 6 commodities are selected. They are coal, natural gas, oil, sugar, steel and iron ore, respectively.

Coal is important in commodities, especially energy sector. It is commonly used in electricity generation and factories, especially in heavy industries. Further, its residues are by-products.

Oil is another essential production raw materials in industries. Besides providing energy source for vehicles, they are commonly used in other industrial productions, especially in
chemical manufacturing and plastic, detergent making industries.

In industrial productions and civil engineering, steel and iron are skeleton of the buildings. In addition, the steel and iron are used to make alloys to make daily products such as household electric appliance, vehicles and infrastructures.

Sugar is a representative in food commodity sector. It is used in food production and energy production. As a main character in food production, it is used to present the food commodity.

3.3 Related Studies

Numerous researchers from other publications and institutions had conducted studies against machine learning in price prediction. For example, a team from University of Nottingham Malaysia had tested on several machine learning models in predicting agricultural product price and found LSTM to have performed the best. [22] Another team in Worcester Polytechnic Institute had tried to do cryptocurrency trading and concluded that the perceptron model performed the best. [19] Researchers from Korea had also summarised a few top performing machine learning models and tested on electricity price prediction. [13] Their results and findings will lay the foundation of our work and will be further discussed below.

3.4 Calculation of commodities Prices

In the modern world, basically every commodity are using USD to evaluate the price. However, the US does not produce and consume everything.

Therefore, some resources in commodities are in other nations, like rare earth in China, coal in Indonesia and Australia also steels from Japan and Germany. Even there are many commodities that are mostly produced in the US, the prices are still calculated in the US for simplicity and some foreign reserve purpose.

3.5 Dynamic Changes in Pricing

Due to the advancement of technology, the manufacturing cost will increase, some are interdependent with each other.
There are other external factors affecting the commodity prices, such as the US inflation rate, the world industrial production, the world stock index and the price of crude oil. [9]
4 Introduction

4.1 Background

Machine learning (ML) has been a hot topic in recent years. Many big finance and risk management firms, like J. P. Morgan and Hull Tactical Hull Tactical Asset Allocation had already started utilizing ML algorithms to increase revenues. [21] [20] With the bull market run recently, it is now high time to investigate how to get into the market. [18] With the combination of the two, we devised a project to do commodity price prediction by utilizing novel machine learning techniques.

The study of the commodity is the easiest among the dynamic markets like stock and digital currencies. The reason is it is heavily affected by geopolitics and more predictable, this means if we managed to quantify geopolitical factor it would be computed and modelled by the machine learning models. It can be used for tradings after graduation and gain profit from that. [1]

4.2 Importance of Commodities in Daily Life

Commodities affect us a lot, most of the products cannot be produced without them. Hence, their prices would be important factors affecting the people’s living standard. Moreover, there are many futures and derivatives which based on the change of evaluated prices. In this study, the definition of commodities would be limited to the 4 categories mentioned above. In addition, the relations between different commodities would be indicated, and deduce how it affects our daily life, using the model to make precise and accurate predictions. [5] [9]

4.3 Core of the Study

In this study, not only data analytic is covered. Commodity prices are seriously affected by the political and economic changes. While some of them cannot be quantified but some can be, such as the seasonal changes and stimulus programs. They have certain patterns and repetitive. By linking these events and data, the patterns can be easily observed and analysed. [8]
There are 4 main parts of the study: The first part is to depict the politics and climate changes, using a time-series approach to relate the productions and consumption. The second part is to make price predictions with intervals using machine learning techniques. The third part is to make a web app to simulate trading in reality. The fourth part is to list out the principle of the statistics and mathematics behind the study.

4.4 Theories Behind the Data Analytic

Data Analytic cannot be independent of the theories, such as statistics and mathematics concepts. They would be included in the project as knowledge and theory foundations, such as computational time, GAM, linear regression and correlation; matrix theory; statistical interference. [3]

This part will be a major part in the study, as mentioned before, it acts a guideline for quantification and predictions in machine learning. It involves different aspects in statistics and mathematics, such as linear algebra when talk about linear relationships, ODE and PDE as we may need to do ordinary/partial differential equations respect to time $t$. Markov Chain/Markov Model for interdependent commodities, linear Regression where linear regressions and Markov Chain are key components of the machine learning and GAM Modelling. [11]
5 Prior Work

5.1 How the Results of Previous Works

From the others’ work, some are not disclosing the relations between different commodities. Also, less mathematics are involved. They have mappings and annotations but they are not presenting in details. Our approach is to find out some interdependent factors for the pricing of commodities and then use the algorithm to predict the future price trends. They either have maps or trends, but they may fail to bring any further details. [15] [12]

The IMF prediction paper is providing a simple linear model, and the whole 28 pages are the elaboration of the simple linear model. Further, a comparison of actual price and predicted price is given. Even it is written by IMF, the paper does not contain many data analytic contents and the model is oversimplified. [16]

5.2 What are the Improvements

There are not many research papers focusing on the prediction part. In another word, most of them are just analysing the existing results without any further elaborations. Some are just focusing on agricultural products, while some are not interactive in prediction, and accuracy are need to be improved.

In this Final Year Project, there will be more detailed explanations with words and data and visualization tools. [6]

In addition, in order to make improvement based on the research papers, we will add confidence intervals, interactive user interfaces, mathematics and statistics concepts, etc.
6 Methodology

6.1 Overview

Firstly, a web scraping from the website will be done. Organize the raw data using csv format.

![Figure 1: A simple web scraping code](image)

Once csv files are ready, use Python to make data analysis and visualisation. Observe some geopolitical changes along the desired time frame and see if there are rapid increase or decrease in price of one or more commodities.

In the mean time, the algorithm should be ready for edition, prepare the python platform and start importing the packages inside the common python files. The algorithm includes data implantation and analysis, then it will be predictions.

After making predictions and intervals, a web app simulating the trading processes will be created. With initial capital, the interest and the revenue will be displayed after buying and selling of the commodities.

In the last part of the FYP, some mathematics and statistics concepts will be included and analyse the principle behind the machine learning and the algorithm.

6.2 Web Scraping

There are not many databases available online as csv files. Therefore, web scraping is needed to grasp some data from the web-pages and make them csv files for further processes.
A google chrome app called "Web Scraper" is helpful in grasping the data.

After thorough screenings, we selected a few data sources, namely International Monetary Fund (IMF), World Bank and IndexMundi. IMF holds a lot of the indices for primary commodity prices, including but not limited to categories such as food, metal and energy. [4] Its data can go back as far as the 1980s. IMF and World Bank are also very trustworthy international authorized organizations. IndexMundi provides a clear, summarized list of popular commodities price data within a 30 year time frame, sourced from World Bank.

![Figure 2: A simple web scraping code](image)

With web scraping results come out from the python interface, extract the data to get information in table form, and convert it to the csv files as desired.
When all the information is gathered, the data frame will be merged so that direct comparison will be allowed, including the key "Month", there will be 11 rows for the commodity (Steel is not available yet), the merged data frame can be used to create correlation matrix.

By merging the data frame, a correlation matrix is created and presented in a form of heat-map. There is a strong correlation between oil and coal.

### 6.3 How to use Machine Learning Algorithms to predict the results

In the procedure, the data from 1991 to 2020 will be included. The first step is to analyse the data from 1991 to 2021, then using the result to "predict" the experimental actual 12 months of price, followed by the data fitting and algorithm debugging. In fact, the process is subjected to changes by the actual availability of the data source.
In the preliminary ML models, we split the data into training and testing sets, in the proportion of 80:20 randomly, then predict the data using sklearn package. The experimental results will be compared with observational data and determine how precise and accurate the algorithm it is, which is known as fine tuning procedures.

In the final product of the ML model, we would use the algorithm to predict the trends in the following 12 months. After that, we would use statistics methods to control the confidence intervals, then use data visualization tools, mainly matplotlib to present the data. [6]
6.4 What are the affecting factors

6.4.1 Climate and Seasonality

The climate is a factor as if the nation is too cold then whatever how much the reserve there is, there must be difficulties in digging and extracting the ores and not suitable for farming and growing crops. Then they rely on the others and the prices in these nations would be more expensive than the others. Similar hypothesis would be proved or disproved in the analysis part.

During the summer there will be less usage of electricity, but in the winter, there would be more electricity spent on heating mainly. Therefore, there would be expectations that winter would have more electricity consumption. Moreover, due to seasonal changes, some factories will shut down and hence resulting in less production. [8]

It is important to look at seasonality for patterns in the day of the week, week of month and month of the year. In the study, most of the agricultural products would be affected by the seasonal changes due to the natures of these products.

![Figure 7: Seasonality](image)

The figure shows seasonal price trend generally coincides with the supply/demand of the cash crop – less supply in the spring, more supply in the fall after harvest. Knowing this seasonal trend, you could devise an investment strategy to profit from it.
6.4.2 Geopolitics

While under COVID-19 and global inflation brought by the quantitative easing, much amount of hot money flowing in and out, boosting prices to increase.

Further, the trade conflicts will occasionally happen in the world. Recently the embargo of Australian coal from China would result in under supply of China, while the other nations would buy more proportions from Australia, which result in a global increase in price of Australian coal.

Recently, an African country Guinea was having a revolution. Therefore, suddenly the supply of Aluminum to China broke. Not only that, the global aluminum prices are increasing because of a sudden drop of supply. Hence, ore distributions and its geopolitical status would be within the scope of the Final Year Project, especially in prediction part. [15]

6.4.3 Summary

In this chapter, the relationships between commodities and other aspects are disclosed. The illustrations would be done by words above. However, there would be multiple perspective from different countries due to complicated international relations.

6.5 Relationships of different commodities

For example, the increased pricing in coal may result in increase in prices for steel and electricity. While making steel, more electricity would be consumed, so that it may fit into a model like \( y = ax_1 + bx_2 \), where \( y \) indicates the prices of steel and \( x_1 \), \( x_2 \) are the prices of electricity and coal. Therefore, some numerical methods to evaluate and determine the coefficient of \( a \) and \( b \). [2]

This is a very simple example, there would be more hypotheses to be included in the analysis part. So, it will be in matrix form and involve some changes, and hence Matlab can be used as an extra tool to solve some differential equations and markov chain problems.
6.6 Time Frame

The time frame is important as it involves something that happened in the past, now and future. Analysing the past and now are not difficult, but to predict the future is a challenge, so my method is to use 6 months to observe and fine-tune.

The model maybe useful even after graduation, and so the leverage and future would be acting as an extension of this Final Year Project here.

The frequency for refreshing the prices can be one month or one day. The frequency would be higher so that there would be more precise results.

6.7 ML models to be tested

There are numerous machine learning methods, but the top 4 models after investigation are graph neural networks (GNN), generative advanced neural networks (GANs), semi-supervised learning and perceptron models. [22] [19] [13] In fact the graph neural network is omitted here because GNN is more suitable for handling graph data structures. As for the generative adversarial neural network, it takes into account the historical data and develops its own synthetic data for the future that mimic the distribution of the training dataset. This model is useful because it does exactly the goal of this project - using historical data prices to produce predictions. The third one is a semi-supervised technique. By providing labeled dataset and unlabeled dataset, it will identify important features in the labeled dataset and use that information to predict the unlabeled dataset. This method has been done by a Korean team to predict electricity price’s up and down movement and by some modifications, we can use it to predict the commodities’ price. [13] In addition, it is highly likely to require some form of unsupervised learning, as we would want our program to uncover the underlying relationship or rules within the data that we previously did not know. The 4th model is a multilayer perceptron. It is a model composed of multiple layers of perceptrons. Each perceptron unit acts like a neuron in our brain, and fires the action potential if a certain condition is met. This model was used by a team in Worcester Polytechnic Institute and concluded that this model gave the best result in predicting cryptocurrency price trend. [19] These are the models worth investigating into. We built preliminary models based on some of them and the results would be further discussed below.
6.8 Neural Network

Neural Network

Neural Networks are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

The existence of neural network is to downgrade the difficulties of real-life problems to different simpler problems. With the help of neural network, the algorithm can learn and model the relationships between inputs and outputs that are nonlinear and complex; make generalizations and inferences; reveal hidden relationships, patterns and predictions; and model highly volatile data such as financial time series data. [7]
6.9 Web App Structure

In a research published by University of Nottingham Malaysia, they also did commodities price prediction, but only on the automated agricultural commodity price. [22] However their structure of web app is a great example for us to build on. For the final deliverable of our project, we will be producing a web app with our discovery and results. It would follow the basic structures shown here.

1. Forecasted page to show all the graphs and data
2. Rescalable x-axis
3. Duration of view
4. Type of commodities
5. Forecast period

6.10 Final Visualization Version of Predictions

The final result would be similar to this. There are many good features by chaipredict.com, such as confidence intervals, interactive user interface with enquiries in different periods of time of different commodities.

The figure below shows the trends of the commodity price in a specific period of time. The price is counted in USD per ton, with different confidence levels, from 10% to 98%, depending on the users. In addition, an illustration of time frame is also provided.
Final Prediction

The final prediction shows the current trends of a single commodity in one option, it should be an interactive control panel for users to check the previous commodities price trends, relating two or more different commodities, and slide bars for different confidence intervals, the trend results are marked in orange.
7 Results

7.1 Preliminary ML Model - RNN

7.1.1 Software Specifications

<table>
<thead>
<tr>
<th>Library</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>3.7.11</td>
</tr>
<tr>
<td>sklearn</td>
<td>1.0.1</td>
</tr>
<tr>
<td>keras</td>
<td>2.3.1</td>
</tr>
<tr>
<td>tensorflow</td>
<td>2.0.0</td>
</tr>
</tbody>
</table>

7.1.2 Data Preprocessing

For this initial testing model, the data used here was pulled from IndexMundi in the format of a csv file with 360 rows, which contained the monthly price for the past 30 years.

We first splitted the dataset into a training set of size 348, and a testing set of size 12, such that we could use the past few decades of data to predict one year of time frame.

![Split gold dataset](gold-360test.csv)
![Split gold dataset](gold-360train.csv)

Figure 10: Splitted gold dataset

After that, we modified the shape of the input data by trimming useless rows and columns. Then the dataset’s feature was scaled into [0-1]. The transformation helped regulating the network, by avoiding values too large or too small to be transmitted in the systems and causing problems.

For the final phase of prepossessing, we set up a batch size of 6. For every 6 price values, their corresponding label would be the next consecutive price value, or the start value of the the next row.
<table>
<thead>
<tr>
<th>Month</th>
<th>Price</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Dec-91</td>
<td>361.73</td>
</tr>
<tr>
<td>1</td>
<td>Jan-92</td>
<td>354.45</td>
</tr>
<tr>
<td>2</td>
<td>Feb-92</td>
<td>353.91</td>
</tr>
<tr>
<td>3</td>
<td>Mar-92</td>
<td>344.34</td>
</tr>
<tr>
<td>4</td>
<td>Apr-92</td>
<td>338.62</td>
</tr>
</tbody>
</table>

... ... ... ...

| 343  | 20-Jul | 1846.51 | 6.60% |
| 344  | 20-Aug | 1960.63 | 6.61% |
| 345  | 20-Sep | 1921.92 | -2.37% |
| 346  | 20-Oct | 1900.27 | -1.13% |
| 347  | 20-Nov | 1866.30 | -1.79% |

[348 rows x 3 columns]

```python
In [3]:

# Feature scaling
sc = MinMaxScaler(feature_range=(0,1))
training_set_scaled = sc.fit_transform(training_set)
print(training_set_scaled[0,:])

[0.06169163]
```

Figure 11: Trimmed pandas data input

Figure 12: Code for a [0-1] feature transformation

```python
# Creating a data structure with 6 timesteps and 1 output
X_train = []
y_train = []

for i in range(6,training_set_scaled.size):
    # appending the 6 previous prices to the list for i
    # we need to specify the rows and simply pick the first and only column
    X_train.append(training_set_scaled[i-1:i,0])
    # appending the 6th price to the list for i
    y_train.append(training_set_scaled[i,0])

# transforming pandas lists to numpy arrays required for the RNN
X_train, y_train = np.array(X_train), np.array(y_train)
print(X_train)
```

... ...

```python
[[0.06169163 0.05744806 0.05712534 0.05153718 0.04816913 0.04739132]
 [0.05744806 0.05712534 0.05153718 0.04816913 0.04739132 0.04947593]
 [0.05712534 0.05153718 0.04816913 0.04739132 0.04047593 0.05643847]
 [0.78003562 0.8333129 0.85243059 0.8619544 0.92869113 1.]
 [0.8333129 0.85243059 0.8619544 0.92869113 1. 0.97272488]
 [0.85243059 0.8619544 0.92869113 1. 0.97272488 0.9608292]]
```

Figure 13: Transformation of dataset with a batch size of 6
7.1.3 LSTM Layers

We used 5 LSTM (long short term memory) layers to make one fully connected neural network. Initially, the weights at each node were distributed randomly. With each iteration, the weights were updated accordingly, with a smaller gradients, ultimately reaching a cost minimum. The LSTM layer provided the input signal with an option to forget. For example, by multiplying the signal with 0 would give 0, by multiplying with 1 would pass the signal to the next layer.

![LSTM working principle graph](image)

Figure 14: A LSTM working principle graph. Signals are forgotten at the forget gate.

7.1.4 Prediction Result

The prediction result is as shown in the figure below. The red line is the predicted price, whereas the blue line is the real price. By simple counting, it is clear that the machine did not correctly predict most the up and down trend of the price movement.

![Gold Price Prediction](image)

Figure 15: Gold price prediction result from Dec 2020 to Nov 2021
An important note here is that in the future version of the ML model, there should be some standardized numerical method of evaluation. The possible evaluation metrics include MSE (mean square error), MAE (mean absolute error), R2 score, etc.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – predicted value of $y$

$\bar{y}$ – mean value of $y$

Figure 16: The equations for MAE, MSE and R2
7.2 Other Models in Testing

In January, we built several other models for testing. However, as it is still in the course of making, we are still making improvement on the code and analysis and evaluation are omitted in this part.

By utilizing sklearn’s RandomForestRegressor (RFR) library, we built a second model to predict gold price. This model is an improvement from the previous version, by the addition of other features and a random allocation of testing period. The previous RNN model only used a time series of price to make the prediction. Since we would want the ML model to better fit the commodity price prediction and its price affecting factors, we made used of some indices, e.g. political stability index, as a numerical representation of the price affecting factor. This gave a feasible workaround as the specific features to be fed into the machine for this project. (Note: More features will be included in the final report in April.)

<table>
<thead>
<tr>
<th>Month</th>
<th>Price</th>
<th>Export - Switzerland</th>
<th>Import - UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/1/1996</td>
<td>399.45</td>
<td>1.46</td>
</tr>
<tr>
<td>1</td>
<td>2/1/1996</td>
<td>404.76</td>
<td>1.46</td>
</tr>
<tr>
<td>2</td>
<td>3/1/1996</td>
<td>396.21</td>
<td>1.46</td>
</tr>
<tr>
<td>3</td>
<td>4/1/1996</td>
<td>392.65</td>
<td>1.46</td>
</tr>
<tr>
<td>4</td>
<td>5/1/1996</td>
<td>391.03</td>
<td>1.46</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>295</td>
<td>8/1/2020</td>
<td>1968.63</td>
<td>1.19</td>
</tr>
<tr>
<td>296</td>
<td>9/1/2020</td>
<td>1921.92</td>
<td>1.19</td>
</tr>
<tr>
<td>297</td>
<td>10/1/2020</td>
<td>1900.27</td>
<td>1.19</td>
</tr>
<tr>
<td>298</td>
<td>11/1/2020</td>
<td>1866.30</td>
<td>1.19</td>
</tr>
<tr>
<td>299</td>
<td>12/1/2020</td>
<td>1858.42</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Figure 17: An example of including political stability index in the input for RFR model

The RFR model’s prediction result and accuracy test are as shown below. The green line stands for the original price, whereas the red line is the predicted price. Another improvement of this compared with the previous model, was that it utilized sklearn’s train test split and to randomly assign training set and testing set to determine the overall accuracy, instead of just focusing on the final year.

Other models that are currently in testing phase include: SVR (Support Vector Machines), Gradient Boosting Regressor, Decision Tree Regressor, etc.
Figure 18: RFR model’s prediction result

Figure 19: RFR model’s accuracy test

Figure 20: SVR model’s implementation

Figure 21: SVR model’s prediction graph
7.3 Difficulties

Finding a suitable data source is the most difficult as most of sources with abundant data are quite expensive. Although we identified a few usable data sources, quantifying factors such as climate and politics are also difficult. However, we can take into account for climatic indexes of temperature, rainfall, sunlight, air pressure, but it might be difficult to get the climate data of all different regions in different time period.

Another major factor we would like investigated into in the course of this year project is effect of political news on commodities price change. Yet, it has a similar problem as above. Although some researches would test for the news effect by counting keyword searches, it is quite hard for us to cover all the topics in all different regions. [14] And it would also greatly increase the dimension and complexity of the model as well as the computation time. One of the solution to solve such complexity, would be identify the tops importing and exporting countries and use existing numerical political indices. This greatly reduced the complexity and handling different keyword searches neural network and also gave us an easy way to quantify features to be fed into our existing machine.

7.4 Limitation

As we are using machine learning and predicting commodity’s price, the traditional technical analysis and indicators for stock market will not be investigated here. However there is no telling whether traditional indicators or machine learning models perform better in price prediction. Also, as time is limited, we do not have the time to quantify news or political disputes by screening keywords on the web or social media in the span of the project as suggested by others’ approach. [14] Any influence on the price fluctuations by such factors can only be considered as “noise” in the system. Moreover, as the data from IMF only go back as far as the 1980, it is highly unlikely that we will be able to train the model with older historical data.
7.5 Conclusion

To sum up, we have determined the commodities to be predicted, the price-affecting factors, what platform to scrape the data, the ML models to be tested, the structure and key components of the web app deliverable.

7.6 Future Work

However, more work is needed to be done in the future. We still need to improve data sourcing, do experimental implementation of different ML models. After a prediction is made, several evaluation metrics will be used by comparing the actual price of the product with the price prediction generated by the machine. A t-test can also be used to compare the mean of the samples and thereby concluding whether they are similar in a meaningful way or not. A web app would also be built to showcase our product at the end of year project.

7.7 Optional Pathways

In future, optional features like mock commodities tradings, futures and leverages will be included if time is allowed. It is not within our scope of the study. Therefore, the main focus would be in prediction and its algorithm also the web app functions.
# Timeline

<table>
<thead>
<tr>
<th>Time</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Oct, 2021</td>
<td>Phase I (Project Plan and Website)</td>
</tr>
<tr>
<td>27 Oct - 10 Nov, 2021</td>
<td>Research Period</td>
</tr>
<tr>
<td>10 Nov - 8 Dec, 2021</td>
<td>Collection of historical data (historical prices)</td>
</tr>
<tr>
<td>8 Dec - 5 Jan, 2021</td>
<td>Implementation of the 3 ML models</td>
</tr>
<tr>
<td>10 - 14 Jan, 2021</td>
<td>First Presentation</td>
</tr>
<tr>
<td>23 Jan, 2022</td>
<td>Phase II (Inclusion of other features)</td>
</tr>
<tr>
<td>18 Apr, 2022</td>
<td>Phase III (Final tuning and evaluation)</td>
</tr>
<tr>
<td>19 - 22 Apr, 2022</td>
<td>Final Presentation</td>
</tr>
<tr>
<td>4 May, 2022</td>
<td>Exhibition of Project</td>
</tr>
<tr>
<td>31 May, 2022</td>
<td>Completion of Project</td>
</tr>
</tbody>
</table>
8 Reference

References


