Report Data Analytics: Gold Price Prediction

Heng Vicheka

Deng Sarita

Seng Kimsun

American University of Phnom Penh, Department of Information Technology Management

ITM 270 001 – Data Analysis

Has Sothea

November 24th, 2024

1. Members' names & contributions

- Seng Kimsun: Introduction, Data Preprocessing
- Deng Sarita: Exploratory Data Analysis
- Heng Vicheka: Model Development, Results and Evaluation, Conclusion.

2. Introduction

a. Objective

The goal of this analysis is to predict gold prices using historical data. Accurate gold price forecasting helps investors and analysts make informed decisions in a fluctuating market.

Our ultimate objectives are:

- Create a model to predict gold price in the future
- Use the knowledge learned in class to implement the model

b. Dataset Overview

The dataset used for this analysis contains historical data on gold prices and other factors. Key features include: Date, Price 2 Days Prior, Price 1 Days Prior, Price Today, Price Tomorrow, Price Change Tomorrow, Price Change Ten, Std Dev 10, Twenty Moving Average, Fifty Days Moving Average, 200 Days Moving Average, Monthly Inflation Rate, EFFR Rate, Volume, Treasury Par Yield Month, Treasury Par Yield Two Years, Treasury Par Yield Curve Rates(10 Years), DXY, SP Open, VIX, Crude.

- Date: the date of the day correlating to the data shown
- Price 2 Days Prior: the gold price in "$/ounce" 2 days before the displayed date
- Price 1 Day Prior: the gold price in "$/ounce" 1 day before the displayed date

- Price Today: the corresponding gold price in “$/ounce” to the displayed date
- Price Tomorrow: the gold price in “$/ounce” on the day after the displayed date
- Price Change Tomorrow: ($/ounce) Price Tomorrow – Price Today
- Price Change Ten: ($/ounce) Price 10 Days Prior – Price Today
- Std Dev 10: ($/ounce) The variation of the gold price within in the last 10 days, sum (price last 10 days to price today – mean_10_days_price) / 10
- Twenty Moving Average: The average of gold price within the last 20 days, sum (price last 10 days to price today) / 10
- Fifty Moving Average: The average of gold price within the last 50 days, sum (price last 50 days to price today) / 50
- 200 Moving Average: The average of gold price within the last 200 days, sum (price last 200 days to price today) / 200
- Monthly Inflation Rate: The percentage change in the price level of goods and services within a month, indicating the economy's inflation trend.
- EFFR Rate: The Effective Federal Funds Rate (EFFR), the interest rate at which banks lend balances to each other overnight.
- Volume: The total amount of assets (e.g., gold) or contracts traded during a specific period, often used to gauge market activity.
- Treasury Par Yield Month: The yield rate on a one-month U.S. Treasury security, reflecting short-term interest rates.
- Treasury Par Yield Two Years: The yield rate on a two-year U.S. Treasury security, often used as an indicator of market expectations for interest rates over the medium term.
- Treasury Par Yield Curve Rates(10 Years): The yield rate on a ten-year U.S. Treasury security, commonly viewed as a benchmark for long-term interest rates.
- DXY: The U.S. Dollar Index (DXY), which measures the value of the U.S. dollar relative to a basket of foreign currencies.
- SP Open: The opening price of the Standard & Poor's (S&P 500) stock market index for the corresponding date.
- VIX: The CBOE Volatility Index, also known as the "fear index," which measures the market's expectation of volatility over the next 30 days.
- Crude: The price of crude oil, typically in barrels, serves as an indicator of energy costs and economic activity.

3. Data Preprocessing

a. Data Cleaning, Handling Missing Values & Handling Outliers

I) Data Preparation

- Centering every word and number is each cell and bolding the header
- Fill every missing data with the mean of those data categories (MAR)
- Clear double spaces and correct the wrong spelling in the first row

II) Structuring the data (start from the second row)

- Changing the number format in date column to long date
- Adding scale to Volume column (troy ounces) by using =CONCATENATE () function
- Changing the number formats in Price 2 Days Prior, Price 1 Day Prior, Price Today, Price Tomorow, Price Change Tomorow, Std Dev 10, Twenty Moving Average, Fifty Moving Average, 200 Days Moving Average, DXY, SP Open, and Crude columns to Currency (USD)

The rest of columns are already percentages without percent symbols, therefore, dividing them with one hundred and change the number formats to percentages.

4. Exploratory Data Analysis

a. Descriptive Statistics

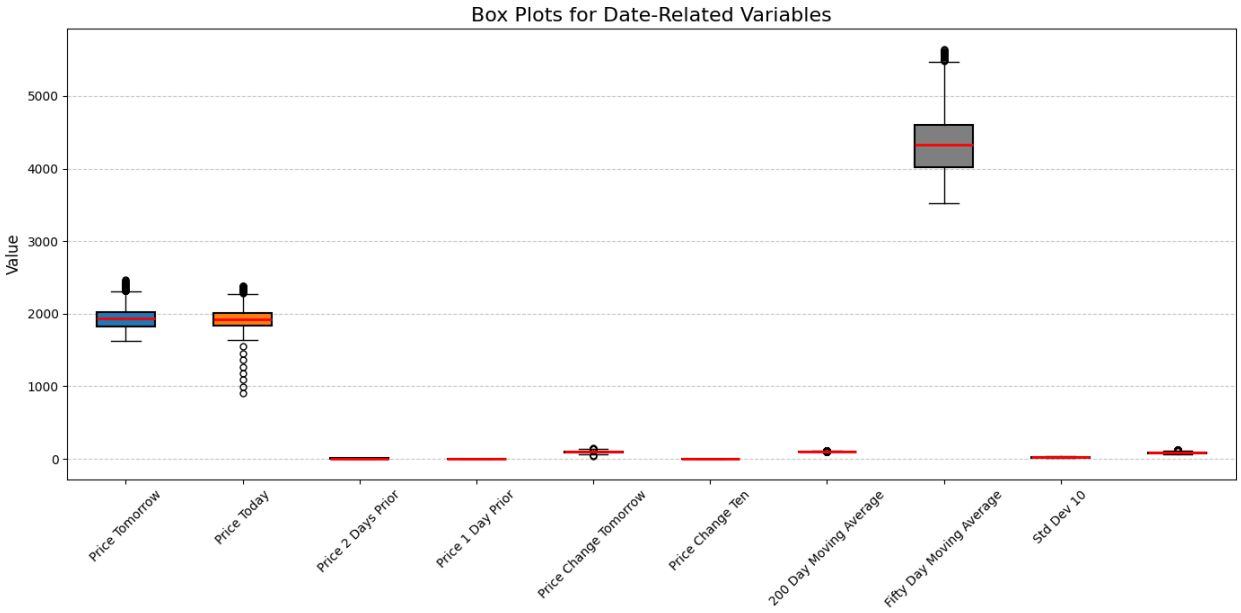
Feature	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Price 2 Days Prior	657	1951.38	186.63	1627.04	1825.18	1930.45	2019.01	2469.65
Price 1 Day Prior	657	1952.27	187.42	1627.04	1825.18	1931.37	2020.19	2469.65
Price Today	657	1953.15	188.16	1627.04	1825.69	1931.44	2022.38	2469.65
Price Tomorrow	657	1953.98	188.78	1627.04	1825.69	1931.50	2022.74	2469.65
Price Change Tomorrow	657	0.83	17.54	-81.89	-9.13	0.66	10.90	66.66
Price Change Ten	657	8.80	55.99	-124.29	-29.79	2.73	42.39	186.17
Std Dev 10	657	20.86	10.48	5.58	13.62	18.15	26.41	62.18
Twenty Moving Average	657	1936.93	195.80	905.14	1832.97	1927.90	2008.97	2387.00
Fifty Day Moving Average	657	1887.04	281.78	362.06	1802.30	1919.30	1992.17	2363.05

200 Day Moving Average	657	1635.71	511.35	90.51	1603.69	1814.88	1949.58	2168.26
Monthly Inflation Rate	657	5.41	2.25	3.00	3.20	4.80	7.90	9.10
EFFR Rate	657	3.70	1.97	0.08	2.33	4.38	5.50	5.30
Volume	657	95.09	15.53	41.00	84.00	96.00	104.00	144.00
Treasury Par Yield Month	657	3.92	1.96	0.02	2.44	4.70	5.51	6.02
Treasury Par Yield Two Year	657	3.66	0.74	1.72	3.28	4.32	4.74	4.98
Treasury Curve Rates (10 Yr)	657	3.34	0.74	1.72	3.28	4.24	4.74	4.98
DXY	657	103.95	3.34	94.87	100.38	104.01	106.04	114.19
SP Open	657	4400.93	490.10	3520.00	4015.54	4327.01	4602.86	5644.09
VIX	657	19.89	6.10	14.19	15.53	19.08	24.07	37.50
Crude	657	84.58	15.35	66.62	76.00	81.27	96.20	124.66

a. Visualization

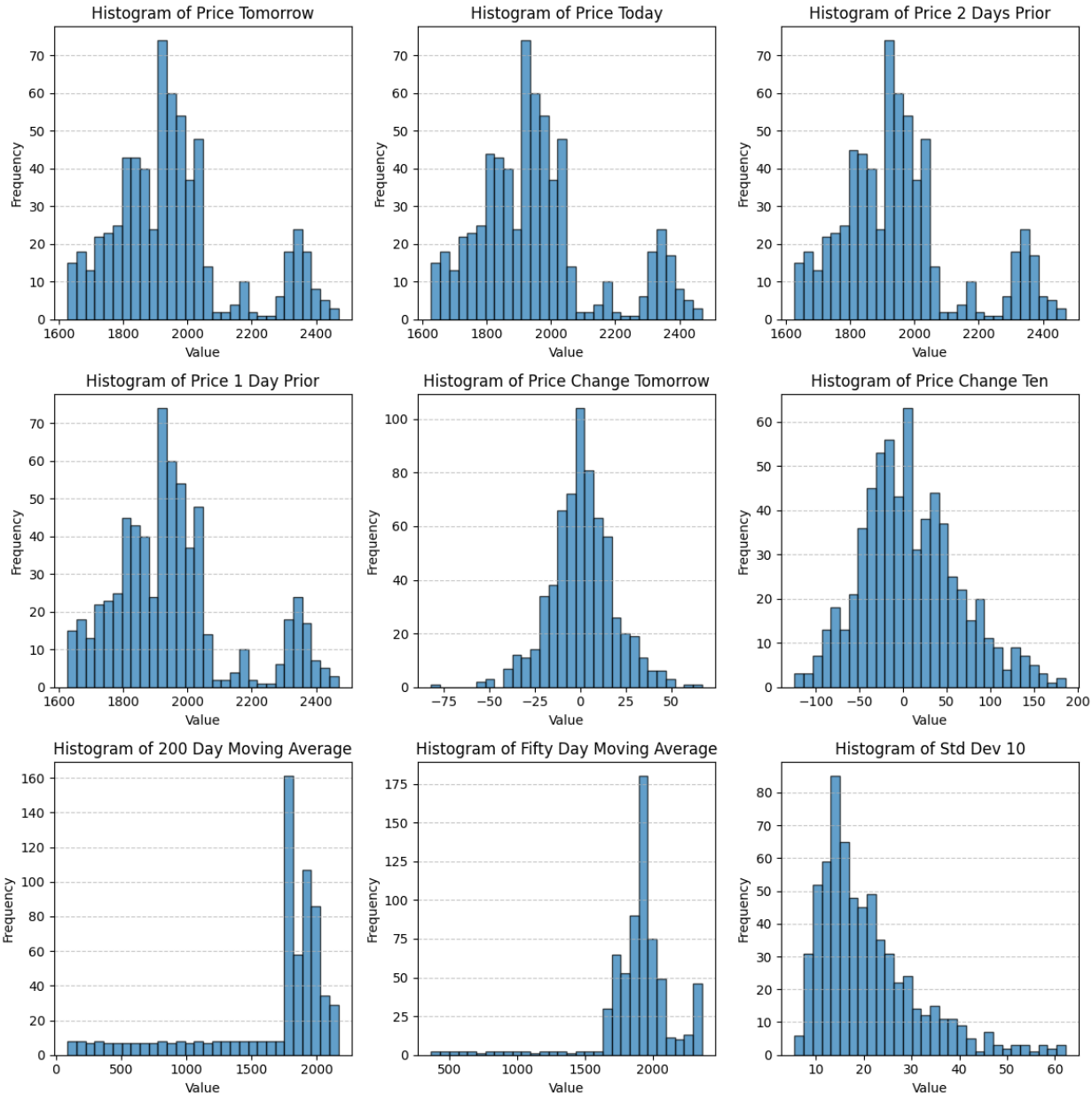
Due to our objective of predicting the future price of gold, and limited resources, we will use only the historical gold prices in this project.

I) Box Plots



Box Plots are used to easily detect outliers and understand the distribution of the data. In our gold price data, the data seems to be in a correct condition with no unusual behaviors, errors, or outliers.

II) Histograms



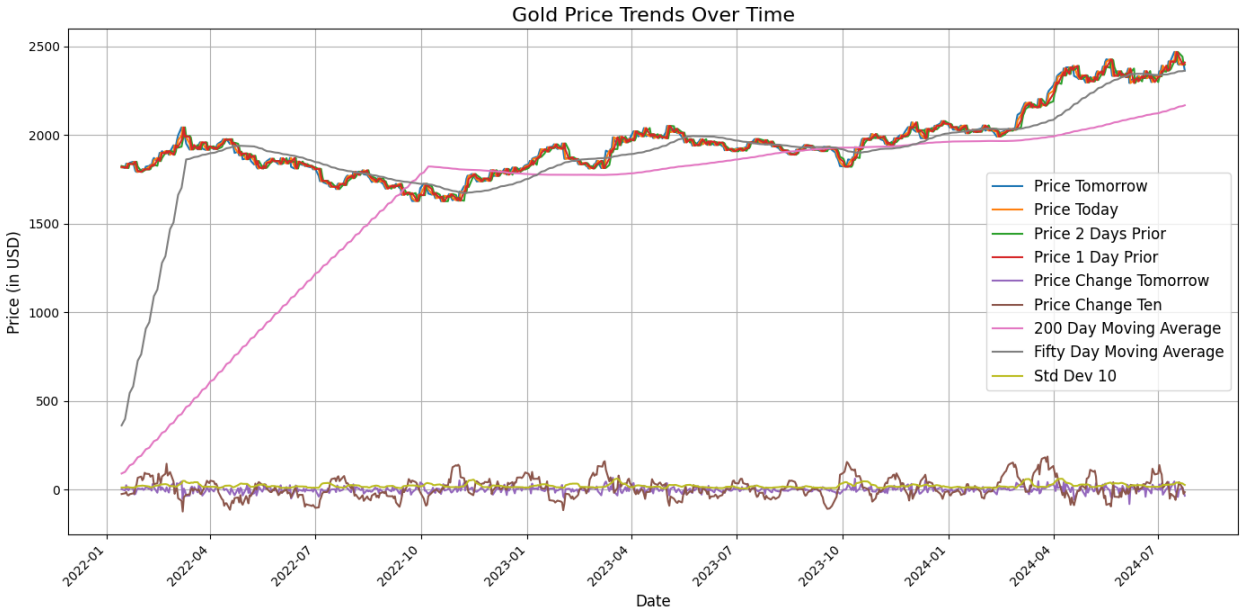
Histograms are used with quantitative data to understand their density and distribution.

Some observations of these graphs are as follows:

- Since Price Today, Price Tomorrow, Price 1 Day Prior, Price 2 Days Prior are technically the repetitions of the same values, they all look almost identical.
- Price Change Tomorrow is centering around 0 in a bell shape, meaning the price usually does not change much between each day.

- The Std Dev 10 is the standard deviations of the gold price for the last ten days. It is skewed to the left showing that within 10 days, the price usually only varies between 10 to 30 dollars, with 60 dollars being the rarest.

III) Trends with Time Plot

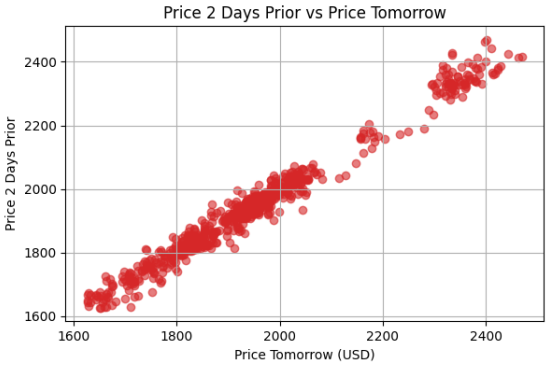


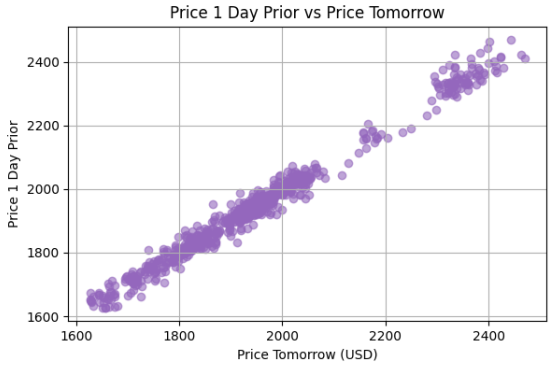
This type of graph is used to visually see the changing trends of data in a quick way.

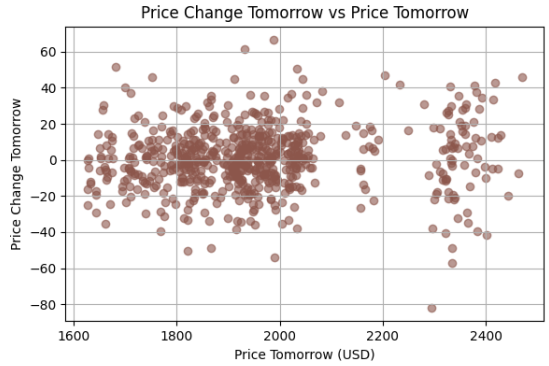
Some observation from this graph is:

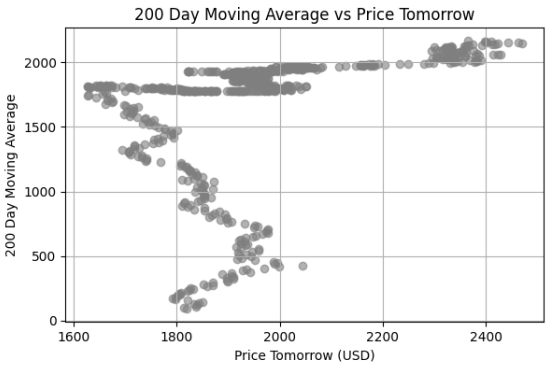
- Gold Prices varied between $1500 to $2500 per ounce from 2022 to July 2024
- Price Change Tomorrow seems to be centering around 0, meaning the price did not change much between each day, which is similar to what we saw in the last histogram charts
- 50 Day Moving Average, the average price for the last 50 days, shows a very useful trend and smooth change. It could make for a great predictor of the future gold price.

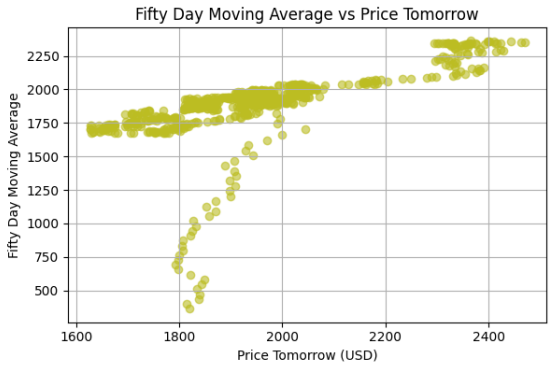
IV)Scatter Plots











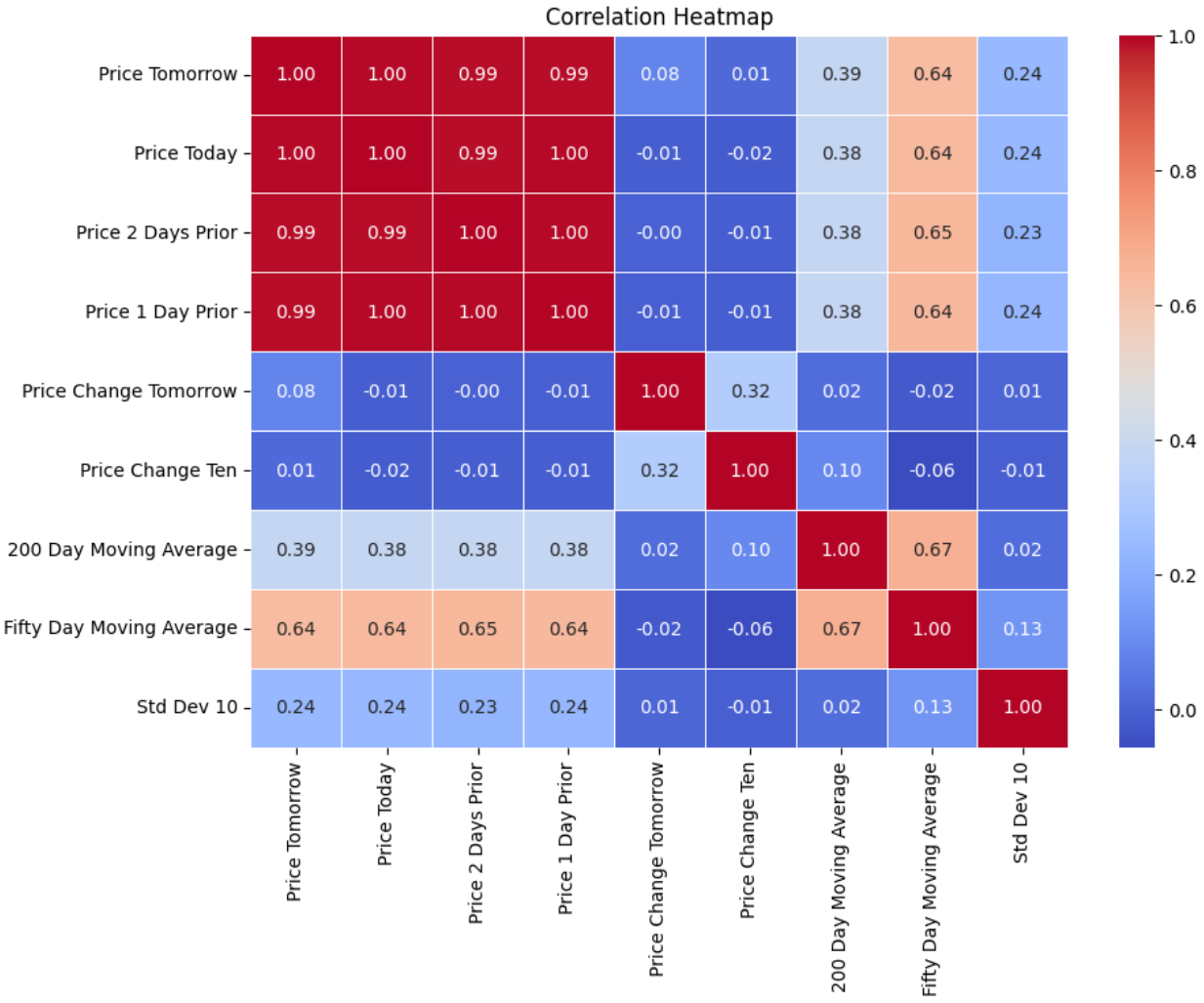
These are used to observe the relationship between variables of data.

The graphs for Price Today vs. Price Tomorrow, Price 1 Day Prior vs. Price Tomorrow, Price 2 Days Prior vs. Price Tomorrow show the perfect positive linear relationship which is expected.

The Price Change Tomorrow vs. Price Tomorrow, and Price Change Ten vs. Price tomorrow shows some potential linear relationships which could be useful for our model. Furthermore, they also show some characteristics of clustering, meaning they could be used to build a Lasso Regression Model. However, due to limited resources and their nature, we are unable to use them in our model.

The last two graphs show non-linear relationships, which we do not want.

V) Heatmap



Heatmaps are exclusively used to see the correlation between variables in an easy way.

Here we can see that the redder, or bluer the cell is, the more correlated the two corresponding variables are. As we have expected, the Price Tomorrow are most correlated to Price Today, Price 2 Days Prior, Price 1 Day Prior are almost or perfectly correlated.

Other potential correlating variables are 50 Day Moving Average and 200 Day Moving Average. However, as we can see in the scatterplots, they do not exhibit great linear relationship at all.

VI) Correlation Table

	Price Tomorrow	Price Today	Price 2 Days Prior	Price 1 Day Prior	Price Change Tomorrow	Price Change Ten	200 Day Moving Average	Fifty Day Moving Average	Std Dev 10
Price Tomorrow	1.000000	0.995674	0.988165	0.991657	0.081228	0.014770	0.385322	0.635212	0.241897
Price Today	0.995674	1.000000	0.991646	0.995672	-0.011729	-0.015088	0.384413	0.639397	0.241915
Price 2 Days Prior	0.988165	0.991646	1.000000	0.995640	-0.002936	-0.014231	0.382256	0.646410	0.228031
Price 1 Day Prior	0.991657	0.995672	0.995640	1.000000	-0.008540	-0.014164	0.383383	0.643117	0.236593
Price Change Tomorrow	0.081228	-0.011729	-0.002936	-0.008540	1.000000	0.320806	0.023166	-0.022778	0.008226
Price Change Ten	0.014770	-0.015088	-0.014231	-0.014164	0.320806	1.000000	0.096192	-0.055840	-0.008404
200 Day Moving Average	0.385322	0.384413	0.382256	0.383383	0.023166	0.096192	1.000000	0.672217	0.023992
Fifty Day Moving Average	0.635212	0.639397	0.646410	0.643117	-0.022778	-0.055840	0.672217	1.000000	0.125716
Std Dev 10	0.241897	0.241915	0.228031	0.236593	0.008226	-0.008404	0.023992	0.125716	1.000000

This table further confirms our observations in the heatmap graph in more detailed mathematical information.

5. Model Building, Training, Evaluating

a) Background

- Inputs: Price Today, Price 1 Day Prior, Price 2 Days Prior, Price Change Ten, Std Dev 10
- Output: Price Tomorrow
- Model Type: Multiple Linear Regression
- Model Main Function: Predict gold price for the day after based on given price today

b) Methods

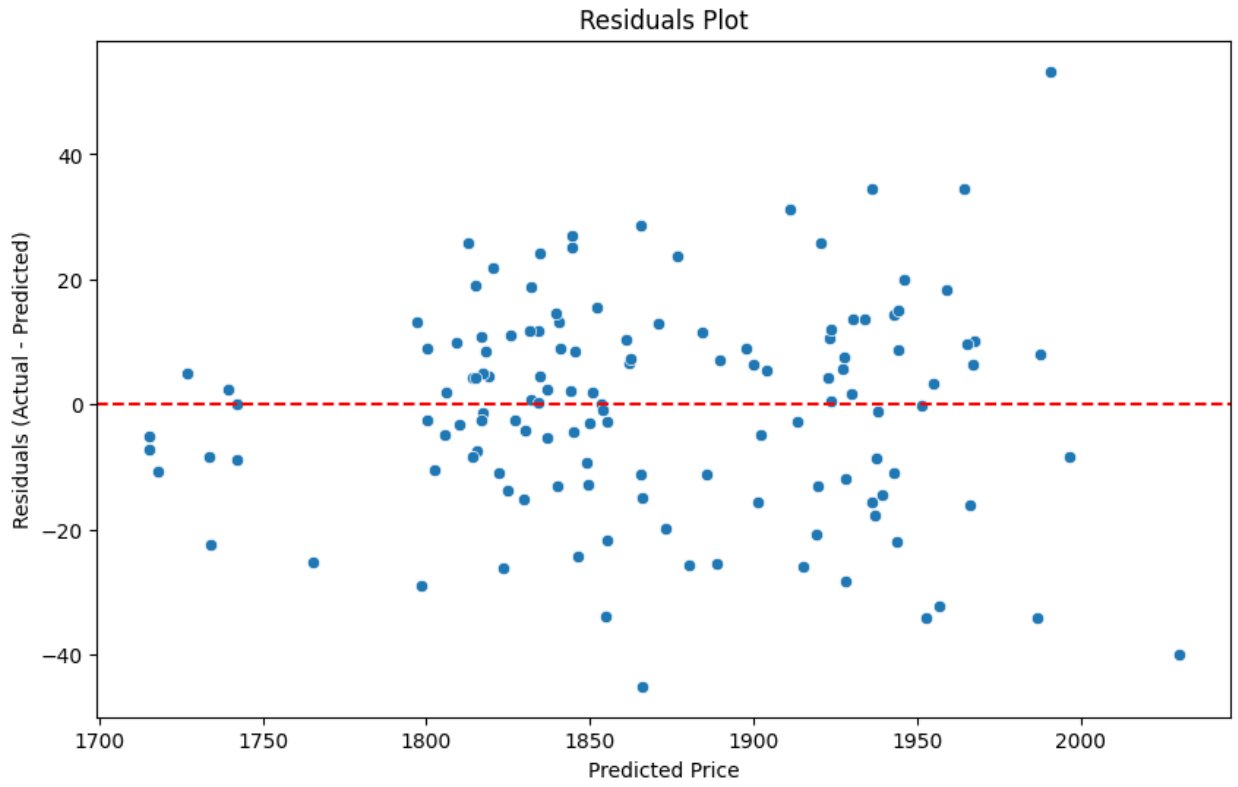
- Create a class for the model with the UML:

GoldPricePredictor
– data: DataFrame – features: List[str] – target: str – model: LinearRegression – scaler: StandardScaler
+ __init__(data, features=None, target='Price Tomorrow') + split_data() -> Tuple + train_model(X_train, y_train) + evaluate_model(X_test, y_test) -> Tuple[float, float, float, float]

+ predict(new_data: DataFrame) -> np.ndarray + plot_residuals(X_test, y_test)

c) Model Evaluating

- MAE (Mean Absolute Error): 13.13
  This shows that the model has the prediction off by mostly $13.13 indicating a good level of accuracy.
- RMSE (Root Mean Squared Error): 16.72
  With an RMSE of 16.72, the model performs well overall.
- R-Squared: 0.9412661699547499
  The model achieved an R-Squared value of 0.941, indicating that approximately 94% of the variance in gold prices is explained by the model's features, which is an excellent result for a predictive model.
- Adjusted R-Squared: 0.9389354624132717
  The adjusted R-squared value of 0.939 is a little less than R-Squared, which is expected; it further supports the model's effectiveness, accounting for the number of predictors used and ensuring that the model isn't overfitting.
- Residuals Plotting:



6. Model Testing on Real Data

- API used to retrieve data today: https://www.goldapi.io/dashboard
- The method used to extract the 5 inputs is: retrieving the price on the given date today or prior and the prices for the previous 10 days and calculating them based on the retrieved information.

7. Conclusion

In this project, we explored the predictive modeling of gold prices using various economic indicators and historical data. The goal was to develop a reliable model that could forecast gold prices with accuracy, focusing specifically on the prediction of the price in the unseen future. By employing the Linear Regression algorithm, we were able to incorporate features such as price 1 day prior, 2 days prior, today, and price changes and standard deviation within the last 10 days.

The model demonstrated strong performance with evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared and Adjusted R-Squared, which collectively showcased the model's ability to make precise predictions. Additionally, the use of standardized features and proper data splitting helped to improve the overall robustness of the model.

Although the result seems decent, the challenges that we have run into made use even more appreciative of it. There were many issues that we had faced, including the lack of field knowledge, data, AI building. Not to mention, the countless Models that we built in the hope of finding a decent one; all of these are due to our lack of experience.

The future scope of this project includes possibly the ability to predict the gold price much longer into the future rather than just 1 day advanced. The model also could use more efficient features with more resources for extra testing and training; instead of relying on historical data, it could be improved by using other economic indicators, which might be better than features used in the model built in this project.

In conclusion, this study highlights the potential of machine learning in financial forecasting, providing a foundation for further exploration and refinement of predictive models in the commodities market. By improving and adapting these models, we can better anticipate market fluctuations and enhance decision-making processes.