GoldPredict: Inflation & Market Impact Insights

HarvardX Data Science Professional Certificate: PH125.9x Capstone 2

Johannes Werner

04 April, 2025

Contents

1 Introduction

The project, GoldPredict: Inflation & Market Impact Insights analyzes and attempts to predict gold prices with the use of inflation rates and market performance by the S&P 500 index. Gold is usually called a safe-haven asset, in uncertain market conditions and its price is influenced by various factors. Understanding the gold price relationship can possibly provide helpful insights for investors.

The dataset includes historical data from the beginning of 2000 to March 2025, including:

- **Gold Prices:** Daily gold prices (GC=F) sourced from Yahoo Finance, as the average of the daily high and low.
- **S&P 500 Index:** Daily S&P 500 index values (^GSPC) sourced from Yahoo Finance, also as the average of the daily high and low.
- **Inflation Rates:** Year-over-year inflation rates derived from the Consumer Price Index (CPIAUCSL) sourced from the Federal Reserve Economic Data (FRED) and calculated as the change in percentage over the previous 12 months.

Sources: ("Yahoo Finance" n.d.), ("Federal Reserve Economic Data" n.d.).

The goal of this project is to predict gold prices by modeling the relationship between gold prices, inflation rates, and the S&P 500 index, using machine learning techniques. The project is divided into the following steps:

- **Data Exploration:** Analyze and visualize the dataset to get an overview and possibly identify patterns and relationships between gold prices, inflation, and market performance.
- **Modeling:** Apply linear regression and random forest models and evaluate their performance using model metrics respectively on a prediction of the gold price the beginning of 2025.
- **Evaluation:** Apply linear regression and random forest models to predict gold prices, evaluating their performance on random time periods.

The report was created using R Markdown in RStudio including the use of the R programming language for statistical computing and data analysis.

2 Data Exploration

This section explores the dataset. The structure of the data is shown and the relationships between gold prices, S&P 500 prices, and inflation rates are visualized. The dataset contains 6,336 rows of information and covers the period from January 3, 2000, to March 12, 2025.

Remark: Due to the fact that inflation data is originally available only monthly, a linear interpolation was applied during data preparation to also get daily values. This ensures consistent data with the gold and S&P 500 data. The interpolation helps to align the dataset for the machine learning models.

The data structure is presented through a summary of key statistics created with the summary command and shows the range of values for each variable (see Table 1).

Table 1: Summary of data metrics

Date	SP500Price	GoldPrice	InflationIndex
Min. :2000-01-03	Min. : 682.9	Min. : 255.6	Min. :-1.959
1st Qu.:2006-04-23	1st Qu.:1202.1	1st Qu.: 578.9	1st Qu.: 1.635
Median :2012-08-04	Median :1498.3	Median :1226.8	Median : 2.231
Mean :2012-08-06	Mean :2139.8	Mean :1148.4	Mean : 2.573
3rd Qu.:2018-11-19	3rd Qu.:2779.7	3rd Qu.:1612.1	3rd Qu.: 3.395
Max. :2025-03-12	Max. :6129.3	Max. :2945.5	Max. : 8.999

In Table 2 exemplarily, the row structure from rows 900 to 905 is illustrated. The format and content of the dataset, including dates and corresponding gold prices, S&P 500 prices, and inflation rates can be seen.

Table 2: Row structure example from line 900 to 905

	Date	SP500Price	GoldPrice	InflationIndex
900	2003-08-04	976.270	348.55	2.207503
901	2003-08-05	973.895	349.00	2.204648
902	2003-08-06	968.290	350.60	2.201794
903	2003-08-07	969.355	351.75	2.198939
904	2003-08-08	977.200	354.75	2.196085
905	2003-08-11	979.835	359.85	2.187522

In Figure 1 an overview of the raw data for gold prices, S&P 500 prices, and inflation rates, plotted separately, can be seen to gain insight into their individual trends over time.

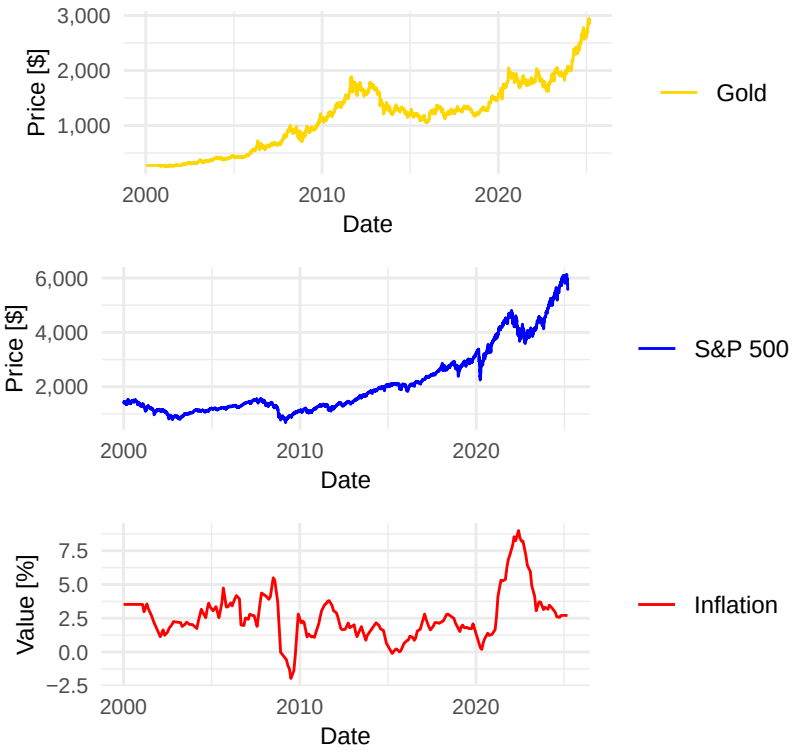

Figure 1: Gold, SP 500 and Inflation chart from 2000 till now

Figure 2 displays the normalized trends of gold prices, S&P 500 prices, and inflation rates from the start of the dataset (2000). Gold price and the S&P 500 are scaled to 100 at the beginning (January 3, 2000) to be able to make a relative comparison. In this period, until March 2025, gold performed best, by showing the greatest relative growth and it was able to preserve value over the crises between 2000 and 2010.

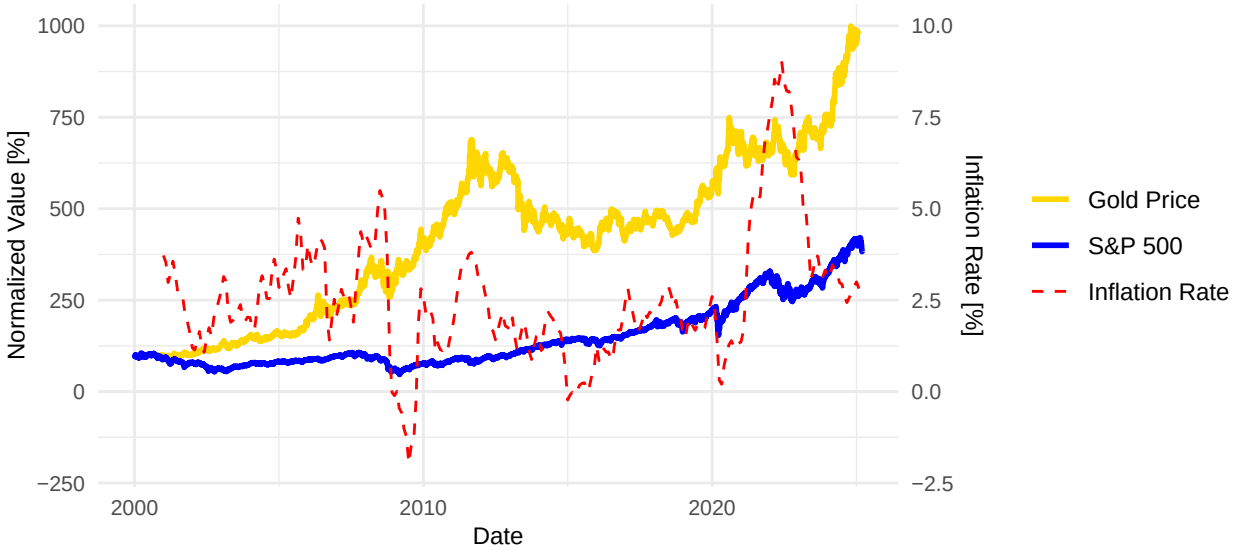

Figure 2: Comparison of Normalized Gold Prices, SP 500, and Inflation Rate from 2000 onwards

In another diagram (Figure 3) the trend starts at 2010. In this case the values are scaled again to 100 but based on the beginning of 2010. In this timeframe, the S&P 500 outperformed gold, which reflects the market's growth and economic recovery.

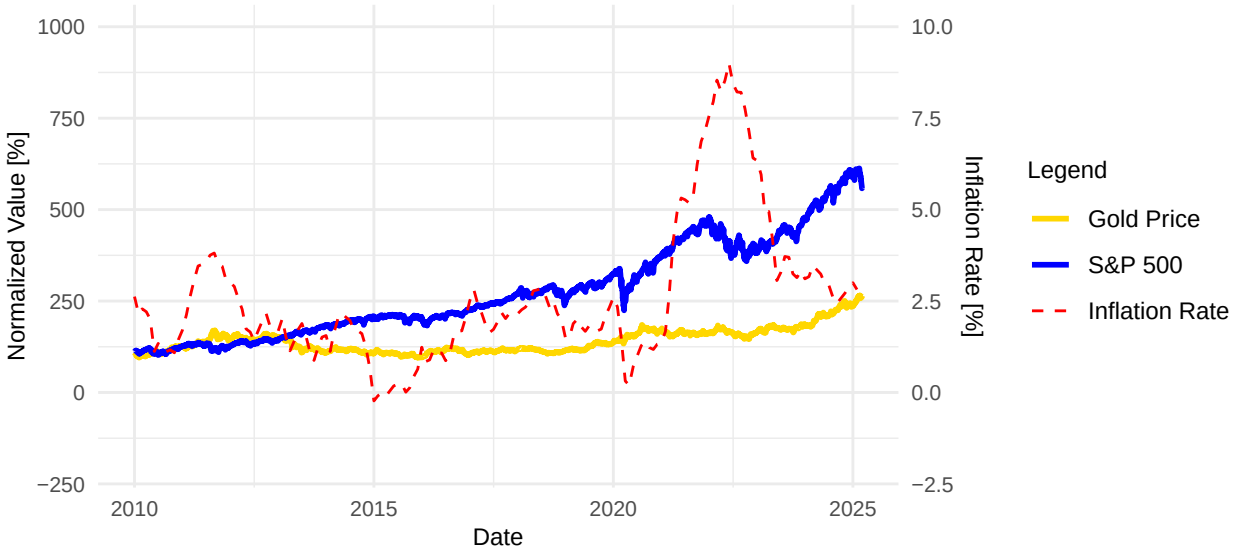

Figure 3: Comparison of Normalized Gold Prices, SP 500, and Inflation Rate from 2010 onwards

So, depending on the chosen timeframe, the gold price relative to the S&P 500 can over long periods develop differently.

3 Methods

In this chapter, the mathematical methods and techniques used for evaluating the gold price prediction models are described. The process is outlined step-by-step, performance metrics are introduced and the model approaches are explained.

3.1 Methodology Overview

The analysis and modeling were carried out as follows:

- Fit of linear model (LM) - check correlation matrix and model residuals to assess its suitability.
- Fit a random forest (RF) model – check performance metrics and variable importance.
- Train the RF model on test data and predict gold prices for the beginning of 2025 – evaluate its performance.
- Train the RF model with cross-validation and compare the results to the model without cross-validation.
- Check the model performance of the RF model on 12 random dates in the data between 2010 and 2025.
- Check the model performance of the LM model on 12 random dates in the data between 2010 and 2025.
- Compare the two models using the performance metrics.

3.2 Performance metric: RMSE

To be able to objectively evaluate the predictive accuracy of the models, two performance metrics were used: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). These metrics are useful for small test sets (small number of prediction days) as they focus on absolute and relative errors and are not variance-based like R-squared, which possibly could mislead with limited data.

Root Mean Squared Error (RMSE): Measures the average squared deviation between predicted and actual values. It is a measure of the average deviation between predicted and actual ratings and defined as (Irizarry 2019, Ch. 33.7.3):

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

Where:

- y_i is the actual gold price
- $\hat{y}_i$ is the predicted gold price
- N is the total number of observations

Mean Absolute Percentage Error (MAPE): Measures the average absolute error as a percentage of the actual value, which leads to a relative measure of prediction accuracy ("Jedox.com" n.d.):

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100$$

Definitions: Same as above.

3.3 Linear model

Linear regression model assumes a linear relationship between the dependent variable (gold price) and the independent ones (S&P 500 price and inflation rate). The model is defined in the following formula (Irizarry 2019, Ch. 31):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Where:

- Y is the dependent variable representing the gold price in USD
- X_1 is the S&P 500 index value in USD
- X_2 is the year-over-year inflation rate in percentage
- β_0 is the intercept of the model
- β_1 and β_2 are the coefficients for the predictors $SP500Price$ and $InflationIndex$
- ϵ is the error term

The model coefficients (β_0, β_1, β_2) are approximated by using the sum of least squares method. Residual diagnostics were performed to check the performance and usefulness of the model.

3.4 Random forest model

Random Forest is an ensemble learning method which constructs decision trees and combines their predictions to get high accuracy and reduced overfitting. Each tree is built on a data sample. At each decision or split a random subset of predictors is taken into account to improve the generalization ("What Is Random Forest? | IBM" 2021) (Irizarry 2019, Ch. 35.2). In this project, the random forest model predicts gold prices (GoldPrice) while its predictors are S&P 500 prices (SP500Price) and inflation rates (InflationIndex). The model has 500 trees ($ntree = 500$) while the variable importance was assessed to understand the contribution of the individual predictors. The final prediction is the average of the predictions from all trees. This makes the model robust to noisy datasets and nonlinear relationships.

3.5 Cross validation

Cross validation is a method to assess the generalization performance of a model by checking it on multiple data subsets. In K-fold cross validation, the data is divided into k folds. The model is trained on $k-1$ folds and the performance is tested on the remaining fold. This procedure is carried out k times, so that every "fold" is tested once. The performance metric is averaged across all folds to estimate the model's performance regarding predictions on unseen data (Irizarry 2019, Ch. 29). In this project, 10-fold cross-validation ($k = 10$) was applied to the random forest model to evaluate the performance and to compare it with a trained model without cross validation. This helps to ensure that the performance of the model is not interpreted as too optimistic due to overfitting on the training data.

4 Results

This chapter shows the findings from the actual analysis and modeling of the gold price with the presented linear regression and random forest models. The results include correlation analysis, model diagnostics, variable importance, predictive performance across different time periods, and comparisons of the models. At the end of the section, the model results are interpreted.

4.1 Correlation Analysis

The correlation plot shows the relationship between gold prices, S&P 500 prices, and inflation rates. It provides insights regarding their interdependencies. In the plot Pearson correlation coefficients are used, shown by numbers, where values range from -1 to 1 (perfect negative to perfect positive correlation) ("Correlation Matrix" n.d.). High positive correlation (e.g., 0.8) between gold prices and S&P 500 prices suggests that as the S&P 500 increases, the gold price tends to rise too, while a negative correlation value might indicate an inverse relationship.

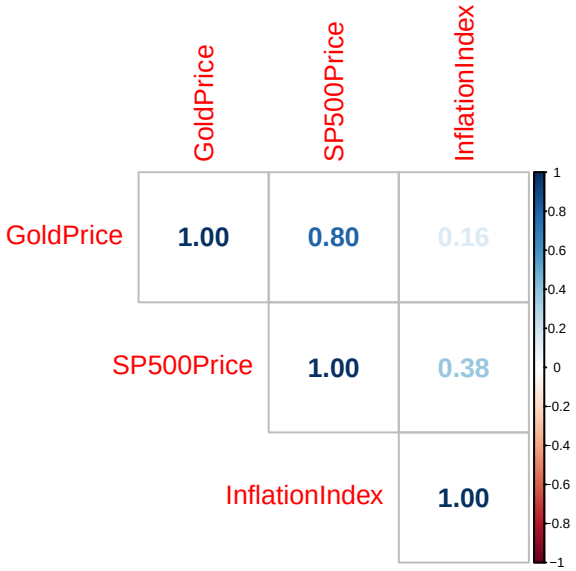

Figure 4: Correlation plot of Gold Prices, SP 500, and Inflation rate

4.2 Linear Regression Diagnostics

The diagnostic plots for the linear model assess if its assumptions — like linearity, normality of residuals, and homoscedasticity — are met. The four plots show: (1) Residuals vs Fitted, to check for nonlinearity; (2) Normal Q-Q, to verify normality of residuals; (3) Scale-Location, to assess homoscedasticity; (4) Residuals vs Leverage, to detect outliers affecting the model.

Deviations in these plots, such as a curved pattern in Residuals vs. Fitted or points far away from the Q-Q line, indicate violations from the assumptions ("Understanding Diagnostic Plots | UVA Library" n.d.).

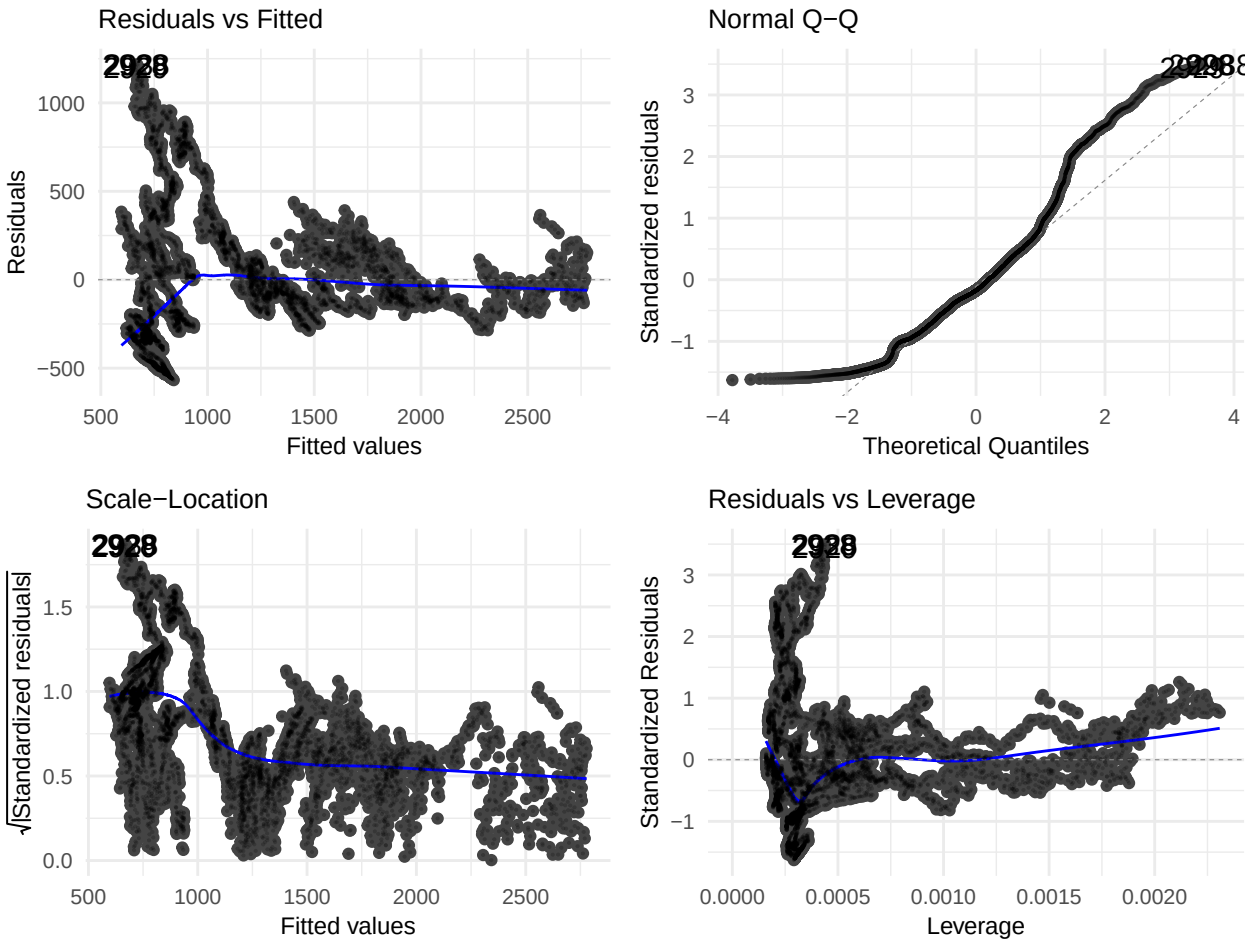

Figure 5: Check of linear regression model

Interpretation of the diagnostics:
("Understanding Diagnostic Plots | UVA Library" n.d.)

1. Residuals vs Fitted

This plot shows a nonlinear pattern, where the residuals are arbitrarily spread around the blue graph. This indicates that the linear model does not accurately capture the relationships in the data. Instead of a random scatter, high positive residuals can be seen at lower fitted values, gradually decreasing. It seems that nonlinear terms in the model might be needed.

2. Normal Q-Q

The Q-Q plot shows the deviations from normality. The middle section follows the reference line relatively well, but there are significant differences at both ends (especially at the upper one). This gives the hint that the residuals have heavier tails than a normal distribution. This could have an effect on the validity of the statistics from the model.

3. Scale-Location

This plot shows heteroscedasticity issues. The spread of the standardized residuals varies across the fitted values, while it seems to be wider at lower fitted values and comes closer at higher values.

The curved blue line confirms this. Equal variance (homoscedasticity) is an important assumption of linear regression. That was violated here.

4. Residuals vs Leverage

The pattern here shows potentially influential observations. Most points lie in the range of lower leverage values but some could have an undue influence on the regression coefficients.

However, none seem to exceed Cook's distance, which suggests that no single observation is critically altering the results.

With these superficial interpretations, overall, these diagnostics suggest that the linear model has several issues that should be addressed: nonlinearity, non-normality of residuals, heteroscedasticity, and potentially influential observations. Alternative model approaches are advised.

This meets the expectation that gold prices are, of course not linear dependent on S&P 500 prices and inflation rates. Because of that, the development of the linear model stopped at this point. Nevertheless, this can be seen as a demonstration on how to possibly do this in a systematic way with more suitable data.

4.3 Random Forest Model Metrics

The random forest model is created by the following code:

```
# Create Random Forest model with 500 trees
rf_model <- randomForest(GoldPrice ~ SP500Price + InflationIndex,
                         data = merged_data,
                         importance = TRUE,
                         ntree = 500)
```

The model's performance metrics provide an initial impression of how it fits the entire dataset.

- **Mean of squared residuals:** 35,894
- **Percentage of variance explained (% Var explained):** 90.22%
- **Number of trees:** 500
- **Number of variables tried at each split:** 1

The metrics indicate that the model explains a relatively high portion of the variance in the gold prices. However, further evaluation is needed to determine whether the mean of squared residuals is acceptable and with that if the model can make good predictions.

4.4 Variable Importance in Random Forest

The variable importance plot in Figure 6 for the random forest model shows the contribution of each predictor (S&P 500 price and inflation rate) to the predictions of the model. Two metrics are shown: the increase in mean squared error (%IncMSE) if a variable is permuted, and the increase in node purity (IncNodePurity) which is based on the variance reduction at each split. The higher the values, the greater the importance. ("RPubs - An R Intro to RandomForest" n.d.) E.g., if

S&P 500 price has a high %IncMSE, exchanging its values significantly reduces the accuracy, which means that it is a key variable of the gold price prediction in the model.

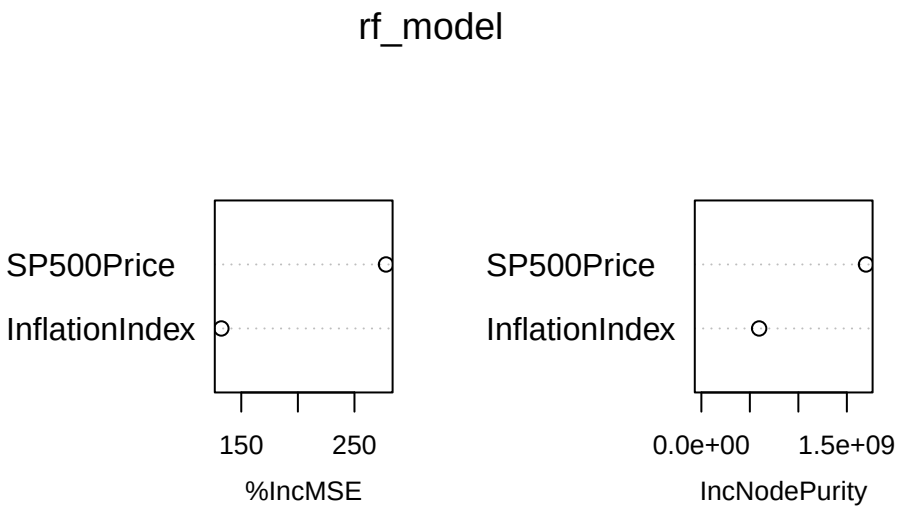

Figure 6: Importance of the variables

4.5 Random Forest Predictions for 2025

The plot in Figure 7 compares the actual gold prices with predictions from the random forest model with and without cross validation. The test period starts in January 2025 and tests the models performance until March 13th. The orange dashed line "Predicted Gold Price (cv)" represents predictions using 10-fold cross validation. The green dashed line shows the predictions without cross validation. As there are no significant differences visible, this suggests that cross validation does not significantly improve the model's performance.

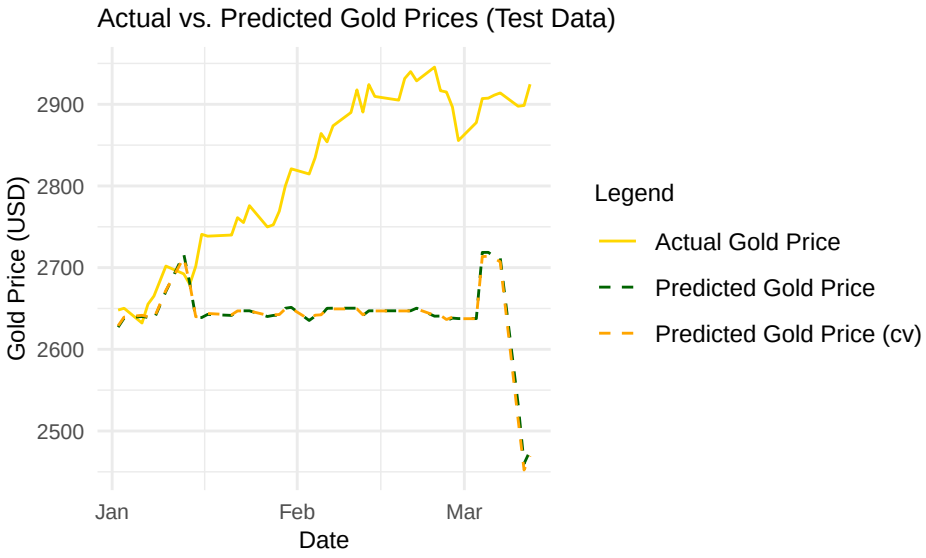

Figure 7: Plot of rf predictions

The first interpretation is that the model performs relatively well for up to two weeks, but afterwards the prediction accuracy seems to disappear. In the next chapter, a series of time intervals is analyzed to evaluate the performance over different periods.

4.6 Random Forest Predictions at Random Time Points

To evaluate the performance across different market conditions, 12 random start dates for prediction between 2010 and 2025 were selected. Only short term is predicted (10 days each) because of the observations in chapter 4.5. The random start dates start at 2010 so that there is at least 10 years before the actual prediction to train the model. The start dates are generated with the following code:

```
set.seed(17) # For reproducibility

# Define date range
start_date <- as.Date("2010-01-01")
end_date <- as.Date("2025-02-28")

# Generate 12 random dates
random_dates <- sample(seq(start_date, end_date, by = "day"), 12)

# Sort the dates for better readability
random_dates <- sort(random_dates)
```

The randomly selected dates are as follows:

```
##  [1] "2010-05-30" "2014-10-24" "2015-05-19" "2015-05-24" "2015-11-26"
##  [6] "2016-12-17" "2020-08-17" "2021-02-22" "2022-05-16" "2024-05-15"
## [11] "2024-09-10" "2024-09-15"
```

With these start dates, the process for evaluating the model involves several steps, which are looped for every date. The code for these steps, the iteration loops, the diagrams and the performance metrics is relatively complicated and long. The basic framework for this process was initially set up manually.

However, the final development and refinement of the code loop was completed with the help of AI tools (cf. Usage of Artificial Intelligence).

The steps in the code are listed below:

- Filter training data up to the start date to ensure the model learns only from historical data.
- Identify the last training data point to anchor the predictions.
- Define a 10-day test period immediately following the last training date.
- Combine the last training point with the test data for continuity.
- Train the random forest model on the training data.
- Generate predictions for the test period and correct them by shifting the first predicted value to the last actual training value to ensure the starting condition.

- Calculate the performance metrics (RMSE, MAPE) to assess prediction accuracy.
- Store the predictions and metrics, and create a plot for each start date showing actual vs. predicted gold prices.
- Combine all plots into a grid for comparison.

The predictions are stored in an extra dataframe with the following structure:

```
##               Date GoldPrice Predicted_GoldPrice  StartDate
## 14759.1 2010-05-28   1207.95           1207.9500 2010-05-30
## 14759.2 2010-06-01   1218.75           1274.9769 2010-05-30
## 14759.3 2010-06-02   1219.50           1020.8885 2010-05-30
## 14759.4 2010-06-03   1212.50            893.0525 2010-05-30
## 14759.5 2010-06-04   1208.65           1137.1852 2010-05-30
## 14759.6 2010-06-07   1227.85           1145.8130 2010-05-30
```

The performance metrics are also stored. This is shown in the following table:

```
##        StartDate      RMSE      MAPE
## 14759 2010-05-30 162.06116 10.416193
## 16367 2014-10-24  72.19085  5.634075
## 16574 2015-05-19  15.26971  1.132712
## 16579 2015-05-24  20.32113  1.691500
## 16765 2015-11-26  69.85580  5.773921
## 17152 2016-12-17  46.04485  3.908830
## 18491 2020-08-17 142.91896  7.313612
## 18680 2021-02-22  48.99175  2.066101
## 19128 2022-05-16  40.56239  2.009249
## 19858 2024-05-15  67.51631  2.484950
## 19976 2024-09-10 147.27142  5.171395
## 19981 2024-09-15 197.82369  7.434368
```

The resulting grid plot in Figure 8 displays the random forest model predictions for each of the 12 random start dates. This allows a visual assessment of the performance across different time periods.

Figure 8: Plot of rf predictions for random start dates

The plot shows more or less random predictions, where no clear or consistent pattern can be seen. This suggests that the model, despite promising metrics during testing, is not suitable for gold price prediction with random starting points. Consequently, it also seems that it should not be used for making predictions about gold price trends for the future.

4.7 Linear Regression Predictions at Random Time Points

Similarly, the performance of the linear regression model is evaluated in the same way as the random forest model to compare their effectiveness. Although the linear regression model was already excluded in testing due to its low performance, it is included here again for comparison purposes because doing this is just a simple code modification.

The resulting grid plot of the predictions for the linear regression model can be seen in Figure 9.

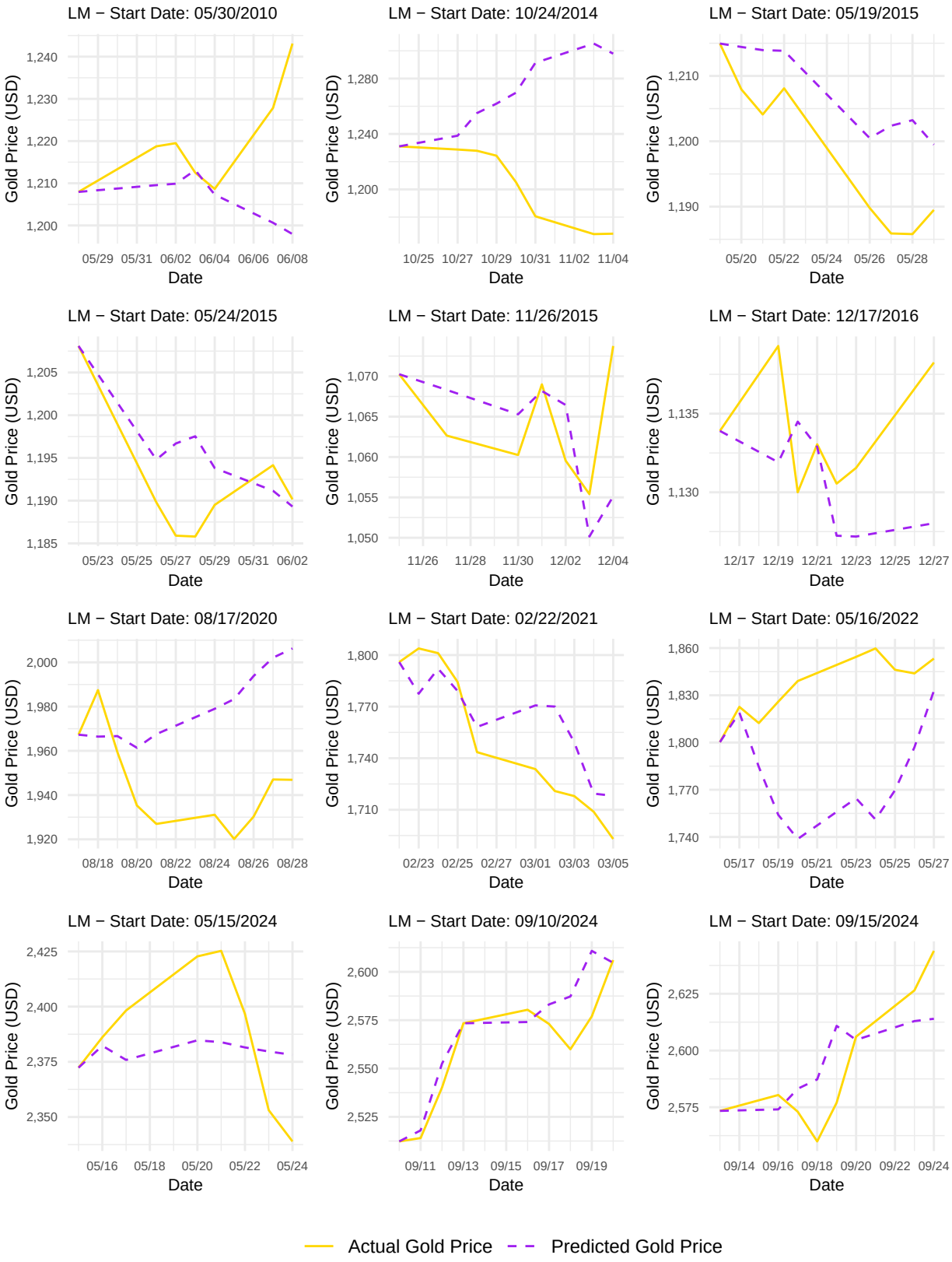

Figure 9: Plot of lm predictions for random start dates

The model aligns relatively well with the actual trends in around 7 out of 12 cases. However, there are time frames where the prediction deviates very much from the observed prices. Overall, the linear regression model demonstrates surprisingly a better prediction performance than the random forest model. But also here, no consistent pattern can be found across the predictions.

The lack of a clear pattern may also not be reliable for predicting future gold price trends, although it seems to be an improvement over the random forest approach.

4.8 Model Comparison

The comparison plot of model accuracy over time contrasts the performance of the linear regression and random forest models using RMSE and MAPE across the 12 random start dates, highlighting differences in predictive accuracy.

The comparison plot in Figure 10 compares the 2 applied models by the defined metrics RMSE and MAPE across the 12 starting points. It shows the differences in the predictive accuracy. Here the visual impression from the chapter before is confirmed. The linear regression model seems to perform better than the random forest model.

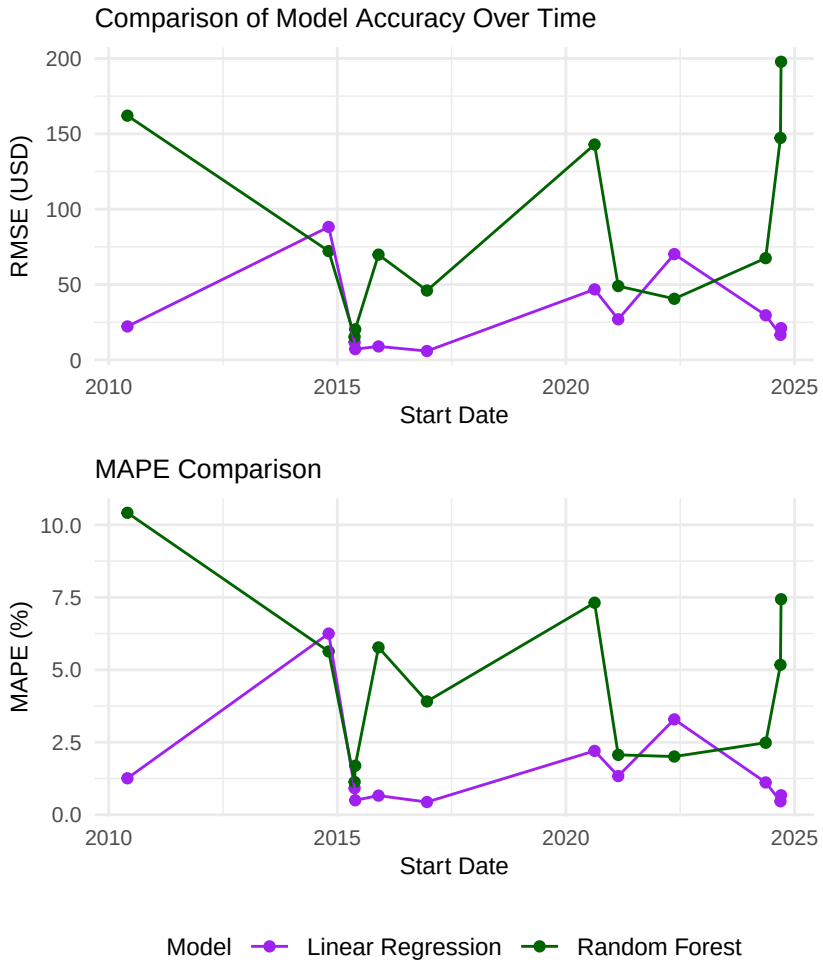

Figure 10: Model evaluation metrics

4.9 Interpretation of Results

The random forest model seemed to be robust in its performance metrics during testing. However, it performed worse than expected when applied to random prediction starting points. Cross validation virtually had no effect on the results. This could be due to the stability of the random forest algorithm in itself.

The linear regression model shows better prediction performance with the actual gold prices in some cases, but both models are not really good for a reliable prediction. This suggests that there are additional factors missing. This could be other financial indices or political and economic influences. These may somehow be implemented in the model to improve the predictive accuracy.

5 Conclusion

In this project, GoldPredict: Inflation & Market Impact Insights, the relationship between gold prices, inflation rates, and the S&P 500 index was analyzed. Predictive models using linear regression and random forest techniques were developed. The models provided insights into short-term gold price movements, but several limitations and areas for improvement were identified.

Key Findings

The random forest model was robust in capturing the nonlinear relationships - but performed in the end worse than expected in the shown scenarios. The linear regression model showed better performance with the actual trends, but not in every time frame or market condition. Both models struggled to make good predictions at random starting points.

The gold price movement is related to Inflation rates and S&P 500 prices, but the impact is varying in different time periods. Other factors, such as geopolitical events or governmental interventions, can have a significant effect on the gold prices, which is not included in the analysis.

Limitations

The dataset includes only inflation rates and S&P 500 prices as predictors. Including broader indices like the MSCI World or macroeconomic indicators could possibly lead to better results. Additionally, political and regulatory interventions have not been considered, which may limit the prediction accuracy during intervention or event phases.

Furthermore, the suitability of the prediction models was not assessed before, which may also have contributed to their limited performance.

Learnings and Future Work

Despite the limitations, a lot of learnings came out of the analysis. Evaluating the model's suitability for the given data beforehand is important. Future efforts in predicting gold prices should focus on identifying patterns in the price movement directly before the prediction periods in such ways that in specific phases the models might perform better. This could lead to being able to find conditions, market phases, or time frames, under which the models align better with the actual trends.

To improve accuracy and applicability:

- Implement additional predictors like interest rates, forex rates, or geopolitical indices.
- Investigate how market conditions directly before the predictions have an influence on the model's performance.

By addressing this, future studies can possibly have enhanced predictive abilities and can provide better insights into gold price movements.

Usage of Artificial Intelligence

Perplexity AI – R code assistance

- **Website:** https://www.perplexity.ai
- **Used in:** Results section (ggplot syntax correction, model testing)
- **Purpose:** Syntax correction and debugging support for ggplot visualizations and assistance in creating the code for data preparation, random dates testing and start value correction for random forest and lm model.

Perplexity AI – Content crosscheck and writing tips

- **Website:** https://www.perplexity.ai
- **Used in:** Introduction and Conclusion sections
- **Purpose:** Crosschecking content accuracy and logical structure; providing suggestions for improvements regarding clarity, precision, and scientific writing style.

AI Grammar Checker | Sapling

- **Website:** https://sapling.ai/grammar-check
- **Used in:** Entire report
- **Purpose:** Grammar and spelling correction.

Grok from X - R code commenting

- **Website:** https://x.com
- **Used in:** R Code
- **Purpose:** Assisting in generating code comments.

References

"Correlation Matrix." n.d. Accessed April 2, 2025. https://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software.

"Federal Reserve Economic Data." n.d. Accessed March 15, 2025. https://fred.stlouisfed.org/.

Irizarry, Rafael A. 2019. *Introduction to Data Science*.

"Jedox.com." n.d. Accessed March 20, 2025. https://www.jedox.com/de/blog/fehlermasze-guete-von-forecasts-ermitteln/#mape.

"RPubs - An R Intro to RandomForest." n.d. Accessed April 3, 2025. https://rpubs.com/mdwybron/487157.

"Understanding Diagnostic Plots | UVA Library." n.d. Accessed April 2, 2025. https://library.virginia.edu/data/articles/diagnostic-plots.

"What Is Random Forest? | IBM." 2021. https://www.ibm.com/think/topics/random-forest.

"Yahoo Finance." n.d. Accessed March 15, 2025. https://finance.yahoo.com/.