Project Report: Analysing and Forecasting Gold Prices

Introduction

The objective of this project is to analyze historical gold prices from January 1950 to July 2020 and develop forecasting models to predict future prices. Gold has been a valuable asset and understanding its price trends can provide insights for investors and economists.

Tools and Libraries

The project employs several Python libraries:

- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical computations.
- **Seaborn & Matplotlib**: For data visualization.
- **Statsmodels**: For time series modeling and analysis.
- **Scikit-learn**: For implementing machine learning models like Linear Regression.

Data Preparation

1. **Data Loading**:
    - The dataset, named gold_monthly_csv.csv, is loaded into a Pandas DataFrame.
    - The dataset contains two columns: **Date** and **Price**, representing the monthly gold prices.
2. **Data Inspection**:
    - The first few rows of the dataset are examined using df.head().
    - The dataset consists of **847 rows** and **2 columns**.
    - The date range of the dataset spans from **January 1950** to **July 2020**.
3. **Data Cleaning**:
    - A new column, **month**, is created using a date range from January 1950 to July 2020.
    - The original **Date** column is dropped, and the **month** column is set as the index of the DataFrame.

Data Exploration

1. **Descriptive Statistics**:
    - A summary of the gold prices is generated using df.describe(), providing insights into the mean, standard deviation, and range of prices.
2. **Data Visualization**:

- A line plot visualizes the trend of gold prices over time.
- A box plot is created to visualize the distribution of gold prices across different years and months.

3. **Time Series Analysis**:
    - The data is resampled to calculate yearly, quarterly, and decade averages, allowing for trend analysis over different time frames.

Modeling and Forecasting

1. **Splitting Data**:
    - The dataset is divided into training (up to 2015) and testing (2016 and beyond) sets to evaluate model performance.
2. **Linear Regression Model**:
    - A linear regression model is fitted using time as the independent variable to predict gold prices.
    - The Mean Absolute Percentage Error (MAPE) is calculated to evaluate the model's accuracy, resulting in a MAPE of **29.76%**.
3. **Naive Forecast Model**:
    - A naive forecasting model is applied, predicting that future prices will remain constant at the last observed price.
    - The MAPE for this model is significantly lower, at **19.38%**.
4. **Exponential Smoothing Model**:
    - The Exponential Smoothing model (additive trend and seasonal components) is fitted to the training data.
    - This model provides a MAPE of **17.24%**, indicating improved accuracy over the previous models.
5. **Forecasting Future Prices**:
    - The Exponential Smoothing model is used to forecast future gold prices for the test set.
    - Confidence intervals for the predictions are calculated and visualized.

Results

- **Linear Regression**: MAPE of **29.76%**
- **Naive Forecast**: MAPE of **19.38%**
- **Exponential Smoothing**: MAPE of **17.24%**

The Exponential Smoothing model shows the best predictive performance among the models evaluated, suggesting that it effectively captures the trends and seasonality in gold price data.

Visualizations

- The following visualizations were created during the analysis:
    - Line plots showing monthly gold prices over time.
    - Box plots for price distributions across years and months.
    - Predictions from the Exponential Smoothing model with confidence intervals.

Conclusion

This project successfully analyzed historical gold prices and developed forecasting models. The Exponential Smoothing model outperformed other models in terms of accuracy, indicating its suitability for time series forecasting in this context. The insights derived from this analysis can be beneficial for investors and stakeholders interested in gold market trends.