Week 11 - Interim Report: Time Series Forecasting for Portfolio Management Optimization

Introduction

This week's project, undertaken for Guide Me in Finance (GMF) Investments, focuses on applying time series analysis and forecasting to enhance portfolio management strategies. The primary objective of this initial phase was to establish a robust data analysis pipeline for key financial assets: Tesla (TSLA), a high-growth stock; Vanguard Total Bond Market ETF (BND), a stable bond fund; and the S&P 500 ETF (SPY), for broad market exposure. Significant efforts were directed towards setting up a structured project environment, fetching historical data via the yfinance API, conducting a thorough exploratory data analysis (EDA) to understand asset behavior, and calculating foundational risk metrics.

Methodology

1. GitHub Repository and Project Setup

The project repository has been meticulously structured to ensure modularity, reproducibility, and maintainability, aligning with best practices for data science projects.

- **Core Structure:** The directory layout includes data/ (for raw and processed data), src/ (for modular Python scripts), notebooks/ (for exploratory analysis and modeling), docs/ (for reports), and images/ (for saved plots).
- **Dependencies:** A requirements.txt file was created at the project root to manage all Python library dependencies, facilitating a consistent environment setup for anyone.
- **Modular Scripts:** Key processes were encapsulated in Python scripts within the src/ directory, including data_fetch.py for acquiring data and preprocess.py for cleaning and merging asset prices.

2. Data Ingestion (YFinance API)

The initial phase involved collecting historical financial data, which serves as the primary dataset for forecasting and portfolio optimization.

- **Script (src/data_fetch.py):** A Python script was developed to connect to the Yahoo Finance (yfinance) API. This script is responsible for downloading historical market data for the specified tickers.
- **Target Assets:** The script was configured to fetch data for **TSLA**, **BND**, and **SPY**.
- **Time Period:** Data was collected for the period from **July 1, 2015, to July 31, 2025**, to ensure a comprehensive historical view.
- **Data Fields Captured:** For each asset, the scraper extracts and stores the Date, Open, High, Low, Close, Adj Close, and Volume.
- **Output:** The collected raw data is saved to individual CSV files within the data/raw/

3. Exploratory Data Analysis (EDA)

To gain a comprehensive understanding of the ingested data's characteristics, an EDA notebook was developed.

- **Notebook (notebooks/EDA.ipynb):** This Jupyter notebook facilitates loading, inspecting, and visualizing the financial data.
- **Key Analyses:**
  - **Data Cleaning:** Checked for missing values and ensured data integrity. The Adj Close price was selected for analysis as it accounts for dividends and splits.
  - **Trend Visualization:** Plotted the adjusted closing prices over time to visually identify long-term trends and compare the growth and volatility profiles of TSLA, BND, and SPY.
  - **Volatility Analysis:** Calculated and plotted daily percentage returns to observe volatility. Further analysis was done using a 30-day rolling standard deviation for TSLA to visualize periods of high and low risk.
  - **Stationarity Testing:** Performed the Augmented Dickey-Fuller (ADF) test on both closing prices and daily returns. This confirmed that prices are non-stationary (requiring differencing for models like ARIMA) while daily returns are stationary.

4. Foundational Risk Analysis

A dedicated section of the EDA was developed to calculate and interpret key risk metrics, providing a quantitative basis for understanding asset risk.

- **Value at Risk (VaR):** Calculated the 1-day 95% VaR for TSLA to estimate the potential maximum loss on a given day. This provides a clear, single-figure risk measure.
- **Sharpe Ratio:** Computed the annualized Sharpe Ratio for all three assets to assess their historical risk-adjusted returns. This allows for a standardized comparison of performance relative to volatility.
- **Outlier Detection:** Identified days with unusually high or low returns (defined as >3 standard deviations from the mean) to understand the magnitude and frequency of extreme market movements for TSLA.

Challenges & Solutions

- **Challenge: Package Installation on Windows:**
  - During the environment setup, the installation of the pmdarima library failed. The error message indicated that Microsoft Visual C++ 14.0 or greater was required to build the underlying C extensions for its dependencies (ecos).
  - **Solution:** The issue was resolved by downloading and installing the "Microsoft C++ Build Tools" from the official Visual Studio website. This provided the necessary compiler for pip to successfully build and install the package wheels.

Future Plan

Building upon the established data analysis foundation, the next steps will focus on developing predictive models and using their outputs for portfolio construction and validation.

1. **Time Series Modeling (notebooks/Modeling.ipynb):**
   - Develop and compare at least two forecasting models (ARIMA and LSTM) to predict future TSLA stock prices.
   - Optimize model parameters and evaluate their performance using metrics like MAE, RMSE, and MAPE to select the best-performing model.
2. **Future Trend Forecasting (notebooks/Modeling.ipynb):**
   - Use the selected model to generate a 6-12 month forecast for TSLA's stock price, including an analysis of the forecast's trend and uncertainty.
3. **Portfolio Optimization (notebooks/Backtesting.ipynb):**
   - Use the forecast for TSLA and historical data for BND and SPY to construct an expected returns vector.
   - Generate the Efficient Frontier and identify key portfolios, such as the Maximum Sharpe Ratio and Minimum Volatility portfolios.
   - Recommend an optimal portfolio and justify the selection.
4. **Strategy Backtesting (notebooks/Backtesting.ipynb):**
   - Define a backtesting period and a benchmark portfolio (e.g., 60% SPY / 40% BND).
   - Simulate the performance of the recommended strategy against the benchmark on historical data.
   - Analyze and compare the cumulative returns and Sharpe ratios to validate the model-driven approach.
5. **Final Report (docs/Investment_Memo.pdf):**
   - Synthesize all findings into a professional "Investment Memo" detailing the methodology, results, and final portfolio recommendation for GMF's investment committee.

Conclusion

This week's efforts have laid a robust groundwork for the portfolio optimization project. A structured repository has been established, data has been successfully ingested and cleaned, and a thorough exploratory analysis has provided critical insights into the assets' historical behavior and risk profiles. This foundation is crucial for the upcoming modeling and backtesting tasks and positions the project well for a successful, data-driven conclusion.