Analysis of Stock market prices of S&P 500 for last 10 years

INTRODUCTION

The stock market is an unpredictable and volatile space, with sudden changes in stock prices having a significant impact on investors' portfolios. As such, forecasting stock market volatility is an essential aspect of risk management and investment decision-making. This project focuses on developing a GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model to predict the volatility of S&P 500 stock market prices. We will explore different variants of GARCH models, evaluate their performance using various statistical measures, and determine the best-fit model for forecasting S&P 500 volatility. Through this project, we aim to provide insights into the underlying dynamics of the stock market and assist investors in making informed decisions.

OBJECTIVE

The objective of this project is to develop a GARCH model to forecast the volatility of S&P 500 stock market prices. We will use GARCH to analyze the data and determine the best-fit model. We aim to evaluate the performance of each model using statistical measures such as Adjusted R2 and F-test, Akaike information criterion, Ljung-Box test, etc.

Whether a GARCH model is a good fit for a particular set of data depends on several factors, such as the nature of the data, the purpose of the analysis, and the performance of the model compared to alternative models. Here are some general criteria that can be used to assess the goodness-of-fit of a GARCH model:

Statistical tests: The model should pass various statistical tests, such as the Ljung-Box test for autocorrelation, the Jarque-Bera test for normality, and the ARCH-LM test for heteroscedasticity. If the model fails these tests, it may indicate that the model does not capture the key features of the data.

Residual analysis: The residuals of the model should be white noise, that is, they should not exhibit any patterns or trends over time. Additionally, the residuals should have constant variance over time, which is the key assumption of GARCH models.

Accuracy of forecasts: The model should be able to generate accurate and reliable forecasts of future values, based on past information. This can be assessed using various measures, such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE).

Economic or financial relevance: The model should be able to capture the economic or financial dynamics that drive the data, and provide useful insights or predictions for the intended purpose. For example, a GARCH model may be used to forecast volatility in financial markets, or to estimate risk measures for investment portfolios.

In general, the goodness-of-fit of a GARCH model should be evaluated in the context of the specific application and goals of the analysis. A good model is one that provides accurate and reliable predictions, captures the key features of the data, and is economically meaningful and interpretable.

DATASET

The S&P 500 dataset is a collection of stock price data for the 500 largest publicly traded companies in the United States. The dataset contains information about the daily stock prices of these companies over the past 10 years, including opening price, closing price, high price, low price, and volume traded. The dataset is commonly used as a benchmark for the overall performance of the U.S. stock market.
The S&P 500 index is calculated using the market capitalization of each company in the index, which is the total value of all outstanding shares of the company's stock. The index is weighted by market capitalization, meaning that larger companies have a greater impact on the index's overall performance.
The dataset includes information on each of the 500 companies that make up the index, including well-known companies such as Apple, Amazon, Facebook, and Google. The dataset is updated daily, providing up-to-date information on the performance of the U.S. stock market.

The dataset can be analyzed using a variety of statistical methods, including time-series analysis and regression analysis. Analysts can use the data to identify trends in the market, make predictions about future market performance, and evaluate the performance of individual stocks or sectors.
The S&P 500 dataset is widely used by investors, financial analysts, and researchers to gain insights into the performance of the U.S. stock market. The dataset can be accessed through a variety of financial data providers and is often used in conjunction with other financial and economic data to inform investment decisions and economic policy.

DATA ANALYSIS

For this study, we used Python to perform the data analysis and modeling. We preprocessed the data using Python's data cleaning and wrangling libraries. We then plotted the open and close prices with respect to date using Python's visualization libraries.

Fig. Date vs Open/Close of F.Y. 2022-23

The present study describes the plotting of high and close price relative to date in order to obtain empirical data for analysis.

Fig. Date vs High/Close of F.Y. 2022-23

This research paper also aims to analyze the open and low price trends of a given dataset with respect to time.

Fig. Date vs Low/Open of F.Y. 2022-23

METHODOLOGY

We calculated the actual volatility, using its formula and with a rolling window of 5 .i.e. number of working days of week, after that we calculated volatility using GARCH and forecasted them. Then we used various tests to figure out the best GARCH model, cross correlation between actual volatility and GARCH volatility etc.

GARCH Model:

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model is a statistical approach used to analyze financial data. It is a time series model that is designed to capture the volatility clustering in financial data, which means that large movements in asset prices tend to be followed by more large movements, and small movements tend to be followed by more small movements.

The GARCH model allows for the conditional variance to be modeled as a function of both past values of the series and past values of the conditional variance itself. This allows the model to capture the persistence of volatility in financial data.

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

- εt is the standardized residual at time t.
- σt is the conditional standard deviation of the return at time t.
- ω, α, and β are parameters to be estimated.

The tests are:

1. LJUNG BOX
2. AIC (Akaike information criterion), BIC (Bayesian Information Criterion)
3. JARQUE BERA
4. HETEROSCEDASTICITY
5. CROSS CORRELATION

1. Ljung Box Test:
   The Ljung Box test is a statistical test used to determine whether a time series is autocorrelated. In other words, it checks if there is a relationship between past and present values of the time series. This test is often used in financial analysis to check whether stock prices follow a random walk or whether there are predictable patterns in the data.

   To perform the Ljung Box test, the time series data is first transformed into a series of residuals. These residuals are then tested for autocorrelation using the Ljung Box test statistic. The test statistic is compared against a critical value based on the degrees of freedom and a chosen level of significance. If the test statistic exceeds the critical value, it indicates that the residuals are significantly auto-correlated and that the time series is not a random walk.

2. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion):
   The AIC and BIC are statistical measures used for model selection. These measures are used to compare the goodness of fit of different models to a given dataset. The AIC and BIC scores can be used to choose the best model from a set of candidate models.

   The AIC is calculated as -2log(L) + 2k, where L is the likelihood function and k is the number of parameters in the model. The BIC is calculated as -2log(L) + klog(n), where n is the sample size. A lower AIC or BIC score indicates a better fit of the model to the data.

   In financial analysis, AIC and BIC can be used to compare different models for predicting stock prices. For example, a model that includes more variables may have a lower AIC or BIC score, indicating a better fit to the data.

3. Jarque Bera Test:
   The Jarque Bera test is a statistical test used to determine whether a sample data set has a normal distribution. It is often used in financial analysis to check whether stock returns or other financial data are normally distributed.

   The test is based on the skewness and kurtosis of the sample data. The skewness measures the asymmetry of the distribution, while kurtosis measures the heaviness of the tails. The Jarque Bera test statistic is calculated as n/6 * (S^2 + 1/4(K-3)^2), where n is the sample size, S is the skewness, and K is the kurtosis. The test statistic is compared against a critical value based on the level of significance. If the test statistic exceeds the critical value, it indicates that the sample data is not normally distributed.

4. Heteroscedasticity:
   Heteroscedasticity is a statistical phenomenon where the variance of the errors in a regression model is not constant across different values of the independent variable. In financial analysis, heteroscedasticity can be a problem when trying to model stock prices or other financial data.

To detect heteroscedasticity, a scatter plot of the residuals can be used to look for a pattern. If the variance of the residuals increases as the predicted value of the dependent variable increases, then heteroscedasticity is present.

There are several methods to correct for heteroscedasticity, including weighted least squares regression and robust standard errors.

5. Cross Correlation:
Cross correlation is a statistical method used to measure the relationship between two time series. In finance, cross correlation can be used to analyze the relationship between two stocks or between a stock and an economic variable.

To perform cross correlation, the two time series are first normalized to have zero mean and unit variance. The cross correlation function is then calculated as the dot product of the two normalized time series at different

We expect to identify the best GARCH model that can accurately forecast the volatility of S&P 500 stock market prices. We will identify the best-fit model among different variants of GARCH and determine its accuracy and reliability. Additionally, we will gain insights into the behavior and dynamics of the stock market over the past ten years.

RESULTS

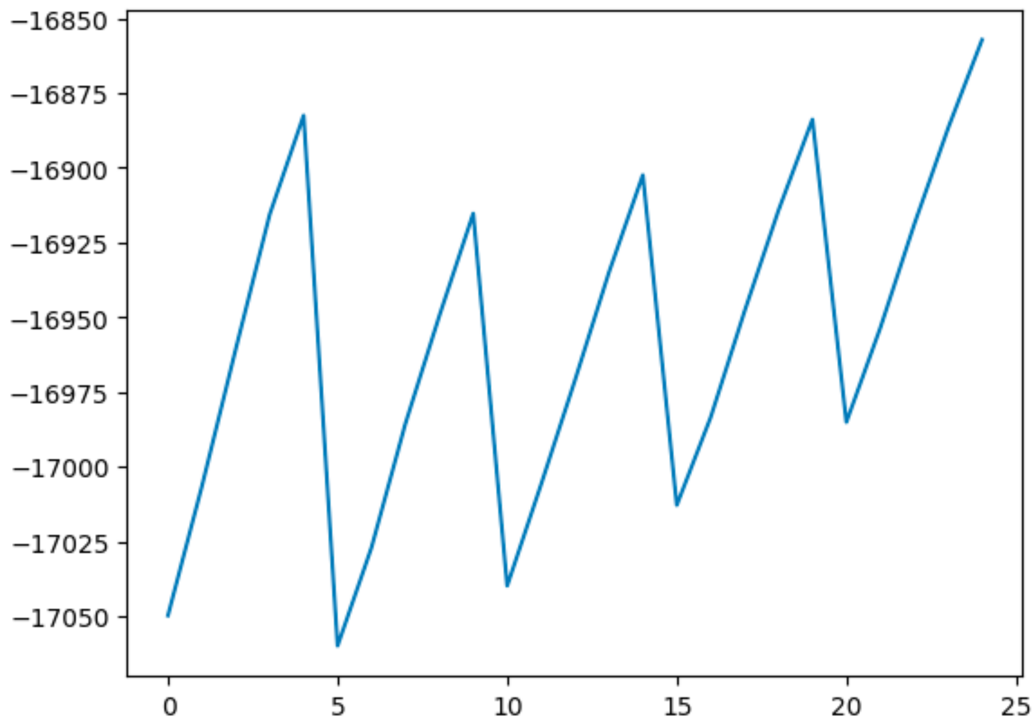
The above-mentioned tests were conducted and they generated the following results:

1. Ljung Box Test: The results of the Ljung-Box test show the LB statistics and p-values for different lags. The LB statistic measures the degree of autocorrelation in the residuals of a model, while the p-value indicates the statistical significance of the test.

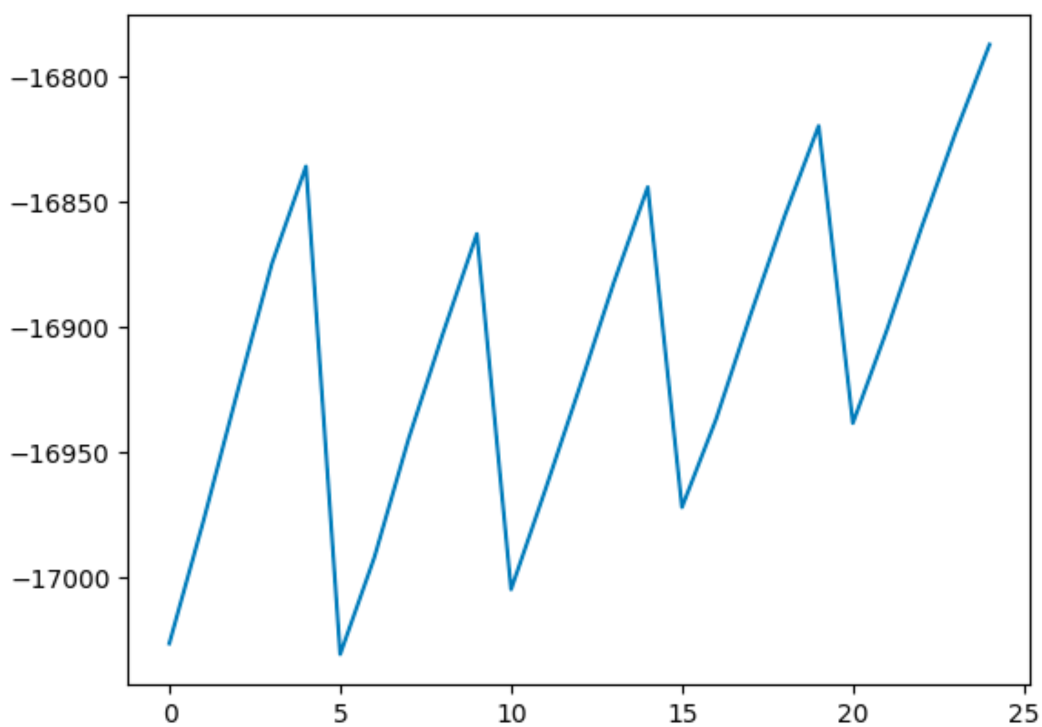
Lags	lb_stat	lb_pvalue
1	49.736124	1.758777e-12
2	70.191153	5.730399e-16
3	70.562995	3.233686e-15
4	76.945698	7.722403e-16
5	83.606342	1.474958e-16
6	115.022002	1.806598e-22
7	171.684024	1.108268e-33
8	215.521938	3.399261e-42
9	264.741066	7.662683e-52
10	274.312172	4.123589e-53

2. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion):

Parameter p	Parameter q	AIC value	BIC value
1	1	-17049.808266838936	-17026.45016451211
1	2	-17006.60403090591	-16977.40640299738
1	3	-16960.590735749363	-16925.55358225913
2	1	-17059.963042949985	-17030.76541504146
2	2	-17027.073008546886	-16992.03585505665
2	3	-16985.59158445888	-16944.71490538695
3	1	-17039.88964620231	-17004.85249271208
3	2	-17005.943840430587	-16965.06716135865
3	3	-16971.02878460011	-16924.31257994647


−16850
−16875
−16900
−16925
−16950
−16975
−17000
−17025
−17050
0
5
10
15
20
25

AIC


−16800
−16850
−16900
−16950
−17000
0
5
10
15
20
25

BIC

3. Jarque Bera Test: The Jarque-Bera test is a statistical test that assesses whether the data have a normal distribution. The test calculates a test statistic and p-value based on the skewness and kurtosis of the data. In this case, the test statistic is 36250.37437102552, and the p-value is 0.0. The very small p-value suggests that the data significantly deviate from a normal distribution.
4. Heteroscedasticity: The results of the Heteroscedasticity test on the residual data of a stock market price dataset provide four values:
   Test statistic - 965.1415547176754.
   P-value - 6.0230211775011075e-201.
   Degrees of freedom - 155.3993868249614.
   Significance level - 3.42092223877066e-254.

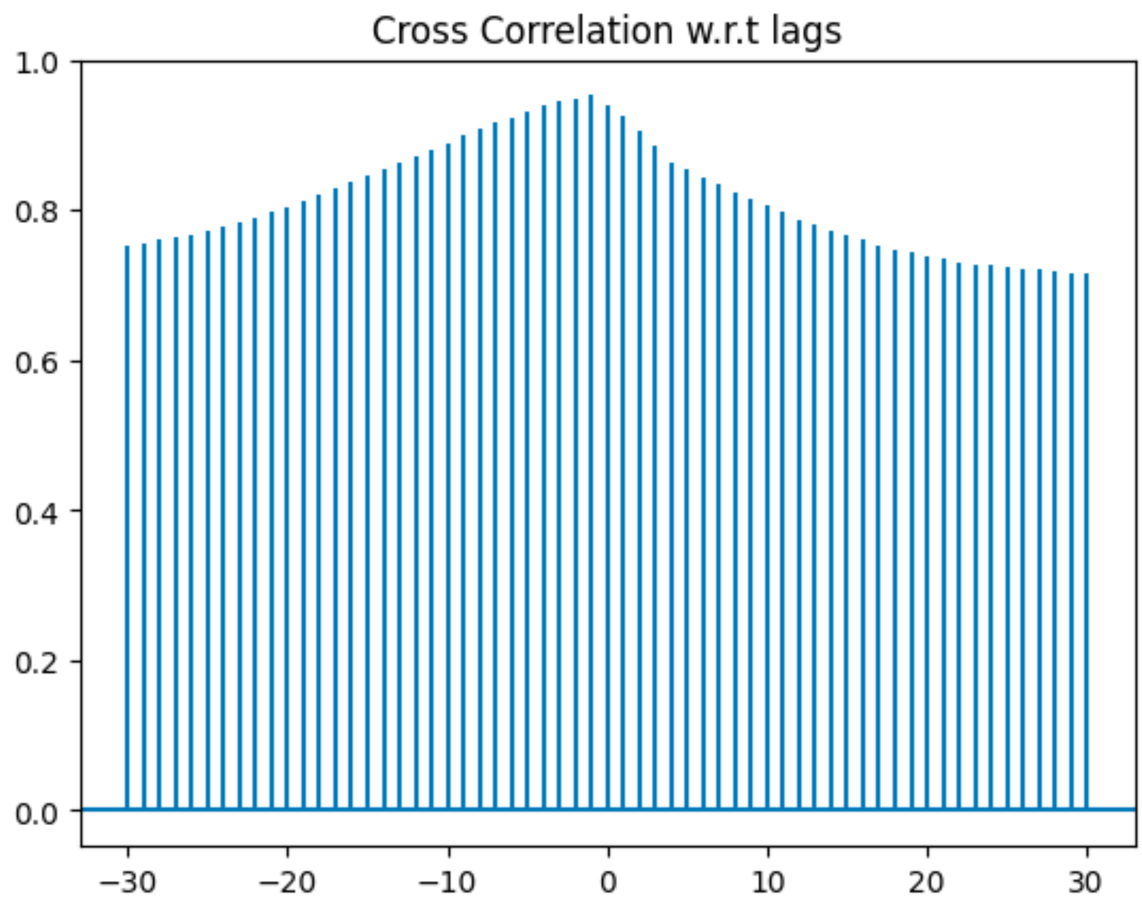
5. Cross Correlation:
   Cross-correlation between 'Close' and 'Open': [-7.37365433e-05 -1.28744925e-04 -2.07855714e-04 ... 9.94198157e-05 4.33896398e-04 1.65527034e-04]. In this case, the cross-correlation coefficients range from -0.000207 to 0.000434, with a few negative coefficients but mostly positive coefficients. The values are very small, indicating a weak linear relationship between the 'Close' and 'Open' prices.

DISCUSSION:

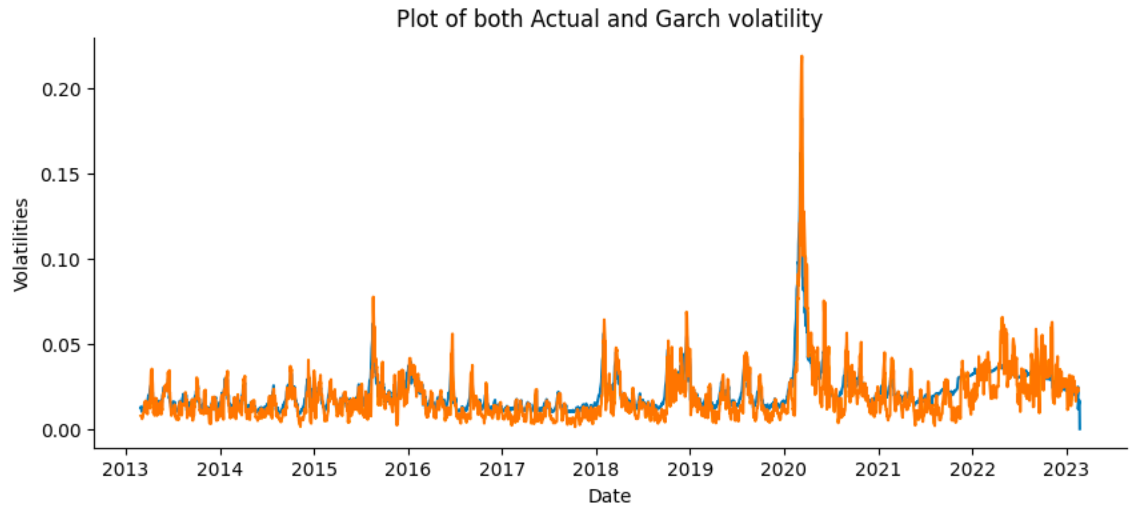
1. Ljung Test
   Based on these results, we can conclude that there is significant evidence of autocorrelation in the residuals of the model, which means that the model may not be adequately capturing all the information in the data. Further investigation may be required to improve the model and reduce the autocorrelation in the residuals.
2. AIC,BIC:
   Based on the value of AIC and BIC, we conclude that the best parameters of p and q are 2 and 1. Hence we use p=2 and q=1 for a best model fit and for other further tests
3. Jarque Bera:based on these results, we can conclude that the stock price dataset being tested does not follow a normal distribution. This may have

implications for any statistical analyses or modeling that assumes a normal distribution, and alternative approaches may be necessary.

4. Heteroscedasticity: based on these results, we can conclude that there is strong evidence of heteroscedasticity in the residual data of the stock market price dataset. This may have implications for any statistical analyses or modeling that assume homoscedasticity, and alternative approaches may be necessary to account for the heteroscedasticity.
5. Cross-Correlation: Therefore, we can conclude that there is little to no significant cross-correlation between the 'Close' and 'Open' prices of the S&P 500 stock price dataset, which means that the 'Close' and 'Open' prices can be treated as independent variables for most practical purposes.



Cross Correlation



Best fit GARCH Model and Actual Volatility

CONCLUSION

Based on the above results, we have come across that the data that the model is fit on isn't very reliable for stock prices prediction since it can lead to various misconceptions and erroneous results.

APPENDIX

– RESEARCH PAPER:

[1] Bhowmik R, Wang S. Stock Market Volatility and Return Analysis: A Systematic Literature Review. Entropy (Basel). 2020 May 4;22(5):522. doi: 10.3390/e22050522. PMID: 33286294; PMCID: PMC7517016.

[2] Baillie, R.T., and R.P. DeGennaro, Stock Returns and Volatility, Journal of Financial and Quantitative Analysis, 25, 203-215, 1990.

[3] Choudhry, T., Stock Market Volatility and the Crash of 1987: Evidence from Six Emerging Markets, Journal of International Money and Finance, 15, 969-981, 1996.

– CODE: In GitHub repository

– DATASET LINK: In GitHub Repository

Team Members Contribution:

Anushree Gupta- started with data analysis and calculated actual volatility.
Rakesh Rathod- used GARCH to forecast the volatilities and performed LJUNG Box test.
Padmasini Krishnan Venkat- Performed AIC, BIC test.
Gautami Mudaliar- Performed JarqueBera and Heteroscedasticity
Abhishek Mathur- Checked Auto and CrossCorrelation