COMPARING FORECASTING ACCURACY USING BAYESIAN GLM, AND BSTS

— Project Report —

Advanced Bayesian Data Analysis

ABDUL MUQSIT FAROOQI

March 17, 2024

TU Dortmund University

Contents

1 Introduction

The accurate forecasting of stock prices holds a growing significance in the vigorous stock market, characterized by volatile returns and risks. Both financial institutions and regulatory bodies have placed considerable emphasis on this aspect. Stocks, as a preferred method of asset allocation, have consistently collected favor among investors due to their potential for high returns. The exploration of stock price prediction has been a persistent area of research, with a historical backdrop featuring the efforts of several economists in the early days to anticipate stock prices.

The motivation for this project originates from the desire to harness advanced statistical techniques to increase the accuracy and reliability of stock price predictions, contributing to the advancing landscape of financial modeling and analysis.

The modeling approach involves the utilization of Bayesian linear regression, and Bayesian structural time series models to predict stock prices. This strategy implies integrating historical stock data and harnessing the uncertainty, and adaptability quantification features inherent in Bayesian methodologies. The primary goal is to catch intricate market dynamics, account for external influences, and establish a resilient framework for forecasting forthcoming stock prices. The models will undergo diligent tuning, validation, and interpretation, ensuring precision and applicability in the dynamical realm of financial markets.

2 Data set and data quality

The data “Comparing Forecasting Accuracy Using Bayesian GLM, and BSTS” is taken from the Kaggle. The dataset contains stock data Volume, Open, High, Low, Date, and Close. The dataset contains 5 independent variables and a dependent variable Close. As the dataset with missing values is already removed therefore no requirement to clean up the data.

Table 1: Variable types and their description.

Variable Name	Variable Type	Description
Company	Categorical	Company of Stock Market
Close	Numeric	Closing price of Stocks
Volume	Numeric	No of shares traded between its daily open and close
Open	Numeric	Starting period of trading
High	Numeric	Highest trading price per day
Low	Numeric	Lowest trading price per day
Date	Date	Date of trading price

3 Models

3.1 Bayesian generalized linear model

Bayesian Generalized Linear Models combine basic GLM with Bayesian methodology and the addition of priors. This is succeeded by the implementation of Markov Chain Monte Carlo (MCMC) sampling. By evaluating Posterior Predictive Checks(PPC) of the Bayesian GLM model the predictive efficacy is determined. This thorough process is also needed the check the credibility of chosen priors and model structure with observed data.

The logarithmic link function for the gamma distribution in the context of generalized linear models (GLMs) is given by:

$$g(\mu) = \log(\mu)$$

where $g(\mu)$ is the link function and μ is the mean parameter of the gamma distribution.

Using the logarithmic link function, the linear predictor (η) is linked to the mean parameter (μ) of the gamma distribution as follows:

$$\eta = \log(\mu)$$

In the GLM framework, the linear predictor η is expressed as a linear combination of the predictors:

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients, and $x_1, \ldots, x_p$ are the predictor variables.**Zhao2021**

The parameters of the gamma distribution are then estimated using maximum likelihood estimation or other suitable methods within the GLM framework.

3.2 Bayesian structural time series model

The Bayesian structural time series model is an advanced approach that integrates Bayesian principles with time series analysis. This model explicitly incorporates temporal dependencies, capturing dynamic patterns and trends. Utilizing probabilistic formulations to account for uncertainties, it offers a robust method for forecasting future values within time series data. Its application extends to various domains, contributing to improved predictive accuracy, especially in financial contexts.**Abdullah2020**

Representation: The local level component assumes that the observed time series is composed of a level component that evolves over time. It captures the underlying trend in the data, allowing for gradual changes.

The local-level model consists of two main equations:

Observation Equation:

$$y(t) = \mu(t) + \epsilon(t)$$

Level Equation:

$$\mu(t+1) = \mu(t) + \eta(t)$$

The level equation signifies that the level at time $t + 1$ is determined by the level at time t plus some random fluctuation.

4 Priors

In this project, priors are used that guide the model based on previous trends and field knowledge. Priors are specifically set, adjusting the average and spread of the data. This approach with clear priors helps the model show our informed views better, making our Bayesian analysis more detailed and fitted to the situation.

4.1 Explicit proper priors

Explicit proper priors involve considering chosen Probability distributions that are assigned to the model parameters.

Parameters such as 'a', 'beta_volume', 'beta_open', 'beta_high', 'beta_low', and 'intercept' are assigned normal distribution. defining these priors with a mean of 5 and a standard deviation of 1. The coefficients for these predictor variables are 5, indicating a moderate effect on the outcome variable.

The parameter 'shape' is assigned a uniform prior between 0 and 50. This allows the parameter shape to be determined by the data provided rather than being strongly influenced by the priors.

The explicit proper priors used have great importance in Bayesian incorporating, modeling, domain knowledge, or informed assumptions about the parameters, guiding the model towards more practical and explicable results.

5 Code

5.1 Libraries

The following libraries are utilized in the code:

- `pandas`: Employed for data manipulation and analysis.
- `numpy`: Essential for numerical operations and array manipulation.
- `matplotlib` and `seaborn`: Utilized for data visualization.
- `pymc3`: A probabilistic programming library for Bayesian modeling.
- `arviz`: A library for exploratory analysis of Bayesian models.
- `StandardScaler` (from `sklearn.preprocessing`): Used for standardizing features.
- `LinearRegression` (from `sklearn.linear_model`): The classical linear regression model for comparison.

5.2 Data loading and preprocessing

The code reads data from a CSV file and filters it to include only the relevant company ('META').

5.3 Bayesian generalized linear regression model

A Bayesian model is set up using PyMC3, defining prior distributions for each parameter, and specifying the linear combination of predictors, and the likelihood function.

5.4 MAP estimation and sampling

The code finds the Maximum A Posteriori (MAP) estimate as an initial starting point for the sampling process. Sampling is performed from the posterior distribution using Markov Chain Monte Carlo (MCMC) with 1000 iterations, 1 chain, and 4 cores.

5.5 Posterior predictive checks and Arviz Conversion

Posterior predictive checks are conducted to assess the model's fit to the observed data. The PyMC3 trace and posterior predictive samples are converted to the ArviZ format for visualization.

5.6 ArviZ Plots

Various plots are generated using ArviZ to visualize the trace of parameters, posterior distributions, and posterior predictive checks.

5.7 Explicit parameter choices

The following explicit parameter choices influence the behavior of the model and are crucial for obtaining meaningful results:

- Number of iterations for sampling: 1000 (`pm.sample(1000, chains=1, cores=4, return_inferencedata=False)`)

- Prior mean for 'a': 0 (`pm.Normal("a", 0, 0.5)`)
- Prior standard deviation for 'a': 0.5 (`pm.Normal("a", 0, 0.5)`)
- Prior mean for other parameters: 5 (`pm.Normal("beta_volume", mu=5, sigma=1)` and similar for other parameters)

These values are selected based on domain knowledge, prior information, or empirical evidence. The number of iterations in sampling determines the length of the Markov Chain and, consequently, the accuracy of the posterior estimation.

6 Convergence diagnostics

6.1 Trace Plots

In Figure 4, the left half of the picture seems steady and doesn't show critical motions.

There are no distinct patterns apparent in these plots; they show up as smooth bends. This suggests that the chains have investigated the boundary space efficiently.

The curves in the plots seem, by all accounts, to be broad as opposed to narrow. They have a smooth and steady top around the value of 5 on the x-axis, bit by bit decreasing in both directions. This broad shape shows a generally wide distribution of parameters.

1. **Graphs Outline**
   - The picture contains seven line charts, each representing different parameters in a Bayesian Generalized Linear Model (GLM).
   - The x-axis spans from 0 to around 800, and the y-axis represents the parameter values (without explicit numerical labels).
2. **Parameter Portrayals**
   - `'a'`: This parameter displays high instability, fluctuating between the values -2 and 2. The posterior distribution of `'a'` shows that the model isn't certain about a particular fixed incentive for the intercept.

- `'beta_volume', 'beta_open', 'beta_high', and 'beta_low'`: These parameters show comparable patterns of variance between values of 2.5 to 7.5. These parameters are highly correlated. They represent the effects of indicators (like volume, open, high, and low) on the close parameter. The posterior distributions show the vulnerability around these effects. The similarity in patterns suggests that these predictors might be correlated or influenced by common factors.
- `'intercept'`: This parameter changes between -3 and 1, with less unpredictability compared to `'a'`.
- `'shape'`: Wavers between around 4 to 4.5. The `'shape'` parameter is defined for the distribution in the model (e.g., Gamma distribution). It determines the shape of the distribution.

6.2 Density plots

In Figure 5, there are density plots, otherwise called Kernel Density Estimations (KDE), representing the distribution of the tested qualities for every boundary in the wake of running the MCMC reproduction.

Plot details

- Each subplot represents a different parameter such as beta_high, beta_volume, beta_open, beta_low, intercept, and shape. The title of each plot demonstrates the boundary name.
- The x-axis shows the scope of sampled values for the boundary.
- The y-axis addresses the probability density.
- Annotations give the mean of the values and the 94% Highest Density Interval (HDI) shows where 94% of the sampled values lie, giving an idea of the uncertainty around the estimate.

Density Curve (Blue Line) This represents the distribution of the sampled values for each parameter. The area under the curve sums to 1.

Mean This is the average of the values for every parameter. It is shown by a vertical line in the plot. It can be seen where a large portion of the sampled values lie (around the mean). If the HDI is wide, it means there's a lot of uncertainty about the parameter's value. If it's narrow, then it's more confident about the parameter's value.

6.3 Posterior predictive checks (PPC) plot

In Figure 1, the Posterior predictive checks (PPC) plot is a method for evaluating the fit of a model by comparing given data with generated data produced by the model. It typically includes actual observed data points, the simulated data points, and often the mean or median of the simulated data.

Need for this Plot PPC is fundamental for model checking. It assists with validating the model by showing if the simulated data can replicate the patterns observed in the actual data. If it is a good fit, then the produced data should overlap the given data.

Results Clarification The plot includes three lines:

- **Observed Data (Solid Line):** Represents the actual data points from the dataset.
- **Posterior Predictive Mean (Dashed Line):** Shows the mean of the simulated data points.
- **Posterior Predictive Samples (Shaded Area):** Represents the range of possible outcomes generated by the model.

Assuming the observed data falls inside the scope of the posterior predictive samples, then the model is capturing the data well and vice versa. It can be seen in the plot that there is a good overlap between the observed and simulated data, suggesting that the model is performing well.

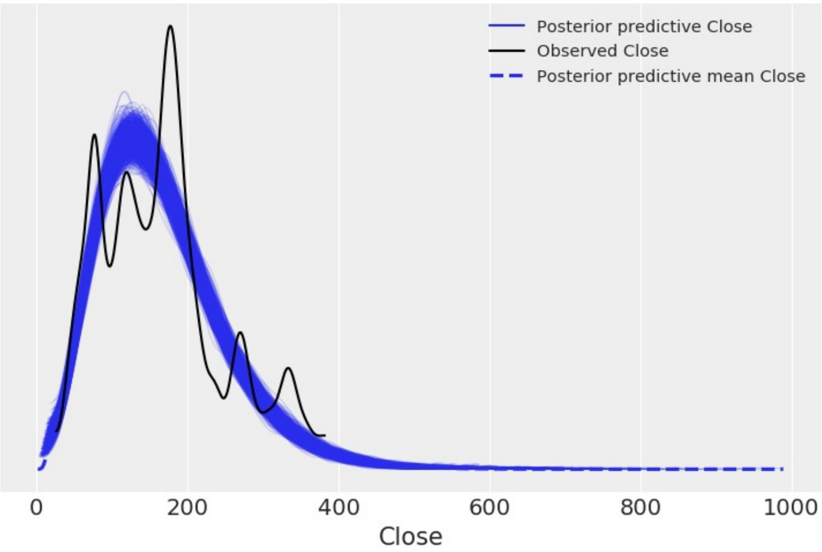

Figure 1: Posterior predictive checks (PPC) plot.

6.4 BSTS local level component - line plot

Model Components: BSTS model comprises a local-level component, representing the underlying trend in the time series data. This component captures the overall level of the data, accommodating fluctuations around this trend. The model aims to encompass both short-term variations and long-term trends.

Forecasted Values: The dashed blue line in Figure 2 signifies the forecasted values generated by BSTS model, extending from the end of the historical data. These predictions are based on observed trends and historical patterns, anticipating future stock prices.

Uncertainty (Confidence Intervals): The shaded orange area around the forecasted values indicates the uncertainty or confidence interval associated with the predictions. Wider intervals imply higher uncertainty, while narrower intervals suggest increased confidence in the forecast. Notably, in the plot, uncertainty tends to increase as the projection extends into the future.

Root Mean Squared Error (RMSE): The Root Mean Squared Error (RMSE) serves as a widely used metric for evaluating the accuracy of regression models. In this context, the RMSE of approximately 57.47 serves as an indicator of how well this Bayesian structural time series (BSTS) model aligns with the test data, specifically closing stock prices. Lower RMSE values correspond to better model performance, suggesting smaller prediction errors.

Visual Assessment: A visual inspection of the plot allows for an assessment of how well the model captures the observed data.

Trend and Fluctuations: The overall trend captured by the model appears to align with the observed data, with the fluctuations around the trend representing short-term variations.

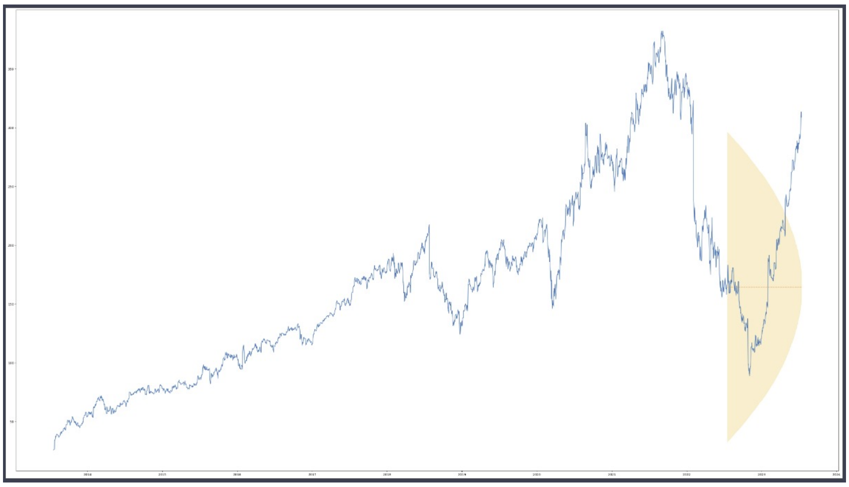

Figure 2: BSTS local level component line plot.

6.5 Level of uncertainty - line plot

Trend Component

Mean Trend: In Figure 3, the solid line represents the average movement of the stock price over time. In this case, it shows an upward trend, suggesting that, on average, the stock price for META has been increasing over the analyzed period.

Uncertainty: In Figure 3, the shaded orange area around the mean indicates the level of uncertainty associated with the trend. It represents the range within which the trend is likely to fluctuate. A wider shaded area indicates higher uncertainty, meaning the trend could vary more widely.

The analysis suggests that while there is an overall upward trend in META's stock price, there can be an additional external factor contributing to variability in the price movement. The uncertainty bands help quantify the level of uncertainty associated with these trends and factors.

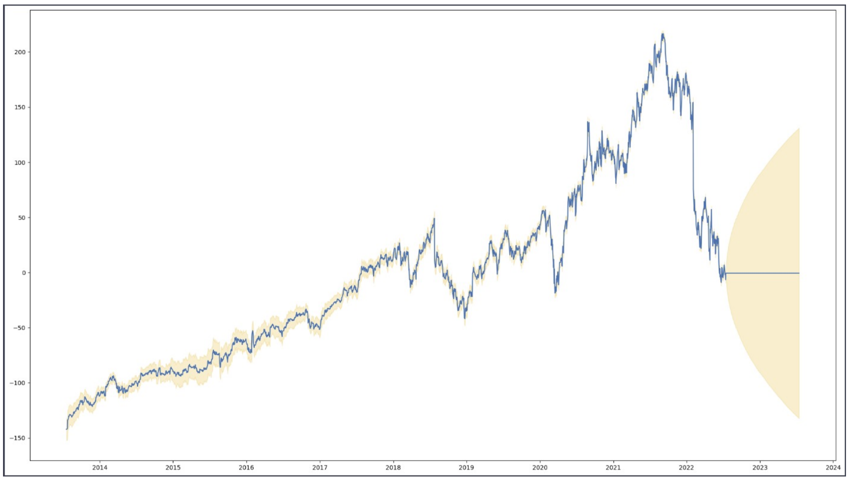

Figure 3: Level of uncertainty - line plot.

7 Model comparison

Bayesian Organized Time Series (BSTS) accomplishes a lower RMSE of 57.47 than the Bayesian Generalized Linear Model (BGLM), which has RMSE of 79.69. Similarly, the mean absolute error of BSTS is 45.22 lower than BGLM of 63.06. This shows that BSTS provides more accurate predictions than BGLM for the considered dataset. Subsequently, BSTS gives off the impression of being a more powerful modeling approach in terms of predictive performance on stock market data.

Table 2: Model comparison.

Error	Bayesian Generalized Linear Model	Bayesian Structural Time Series Model
RMSE	79.6	57.47
MAE	63.06	45.22

8 Limitations and potential improvements

8.1 Bayesian generalized linear model

Limitations

1. Bayesian Generalized linear models assume linear relationships between predictors and the response variable. This might not be true in real-world stock market data, in which the relationships can be non-linear.

2. Generalized linear models might not cope with complex patterns of the variables, potentially leading to underfitting.

3. The prior distributions and model specifications can significantly impact the results. It can be difficult to find the best specifications.

4. Without regularization of Bayesian GLMs, there can be overfitting to the training data, leading to poor generalization performance on testing data.

Potential Improvements

1. Incorporate domain knowledge to engineer new features that better capture the underlying dynamics of the stock market.

2. Use spike-and-slab priors or Bayesian LASSO to automatically select related features for the model and identify the most influential predictors for effectively handling high-dimensional data.

3. Capture temporal dependencies and evolving relationships over time using dynamic models, particularly in financial data where the underlying dynamics may change over economic conditions.

8.2 Bayesian structural time series model

Limitations

1. The computational demands of Bayesian inference techniques can be substantial, particularly with large datasets or complex models.

2. Determining the optimal combination of model components and their parameters is a complex task, given the numerous possible models.

Potential Improvements

1. Efficient handling of calculations may necessitate specialized hardware or software.

2. Selecting suitable prior distributions for model components requires domain knowledge and expertise. The impact of this decision is significant, as inappropriate priors can lead to biased or inaccurate forecasts.

3. Techniques like cross-validation or information criteria may be employed to identify the most appropriate model.

9 Conclusion

The dataset examined in this project comprises information on the closing price of the stock market, along with corresponding information such as Volume, Open, High, and Low. It consists of 2516 observations.

The objective of the Bayesian Generalized Linear Model analysis was to forecast the closing price of the META company that elucidates the relationship between close and various features, such as Volume, Open, High, and Low. The dataset was complete, with no missing values. To check the accuracy of the model the RMSE was calculated which was 79.69.

The objective of the Bayesian structural time series model (BSTS) analysis was to forecast the closing price of META company that elucidates the relationship between close and various features, such as Date. The dataset was complete, with no missing values. To check the accuracy of the model the RMSE was calculated which was 57.47.

Following the models, firstly a Bayesian GLM model was fitted using all variables in the dataset except the date variable. Subsequently, the posterior predictive check was plotted visualizing the posterior predictive close, observed close, and posterior predictive mean close in Figure 1. Finally, to check the accuracy of the model, RMSE was calculated.

Secondly, a Bayesian structural time series model (BSTS) was fitted using close, and date variables in the dataset. Subsequently, the BSTS Local Level Component was plotted visualizing the variance, observed line, and predictive line in Figure 2. Finally, to check the accuracy of the model, RMSE and MAE was calculated.

For future studies, the methodology employed in this project can be extended to a larger scale, encompassing a greater number of observations and feature selection can be used to make studies better. For instance, the model could be trained on a dataset to forecast closing prices, ensuring less variability of the data, and thereby reducing RMSE.

10 Reflection on own learnings

Through this project, we have valuable insights into the principles of Bayesian modeling, time series analysis, and financial forecasting. We have learned how to implement Bayesian GLM, and BSTS models, interpret the results, and assess model performance.

This experience has deepened our understanding of Bayesian inference, prior selection, and the challenges associated with forecasting financial time series data. We have also gained practical skills in data pre-processing, model fitting, and evaluation, which are valuable for future research and professional endeavors in data science and quantitative finance.

11 References

[1] Azam Hajiaghajani (2023). *Designing a combined Markov-Bayesian model in order to predict stock prices in the stock exchange. International Journal of Innovation in Management, Economics and Social Sciences*, 3, 33-41. DOI: 10.59615/ijimes.3.2.33.

J.I.2023

[2] J.I. Eghonghon, Daniel Aliu, Kingsley Ukhurebor, and A.S. Mankide (2023). *A Comparative Analysis of Fuel Price Forecasting in Nigeria Using Bayesian Structural Time Series, Vector Regression, and ARIMA Model*, 8, 13-18.

Zhao2021

[3] Xuejun Zhao (2021). *Application of Bayesian Regression Model in Financial Stock Market Forecasting.* DOI: 10.2991/aebmr.k.210909.019.

Abdullah2020

[4] Abdullah Almarashi and Khushnoor Khan (2020). *Bayesian Structural Time Series. Nanoscience and Nanotechnology Letters*, 12, 54-61. DOI: 10.1166/nnl.2020.3083.

12 Appendix

Additional figures

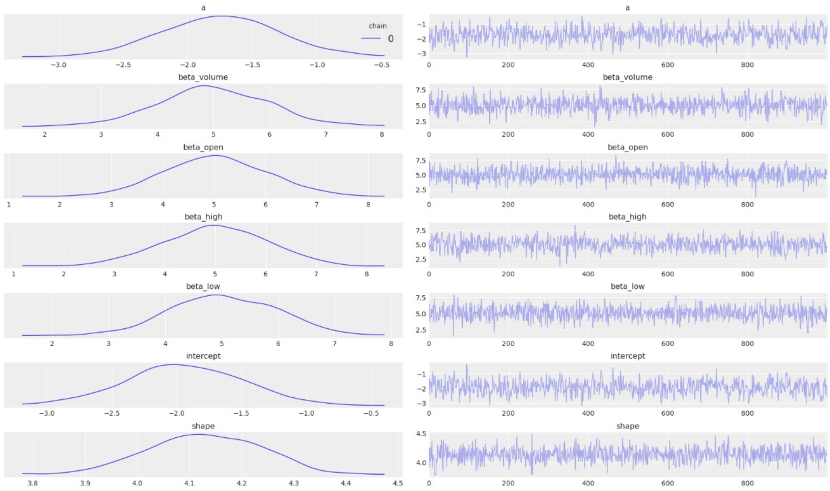

Figure 4: Trace plot.

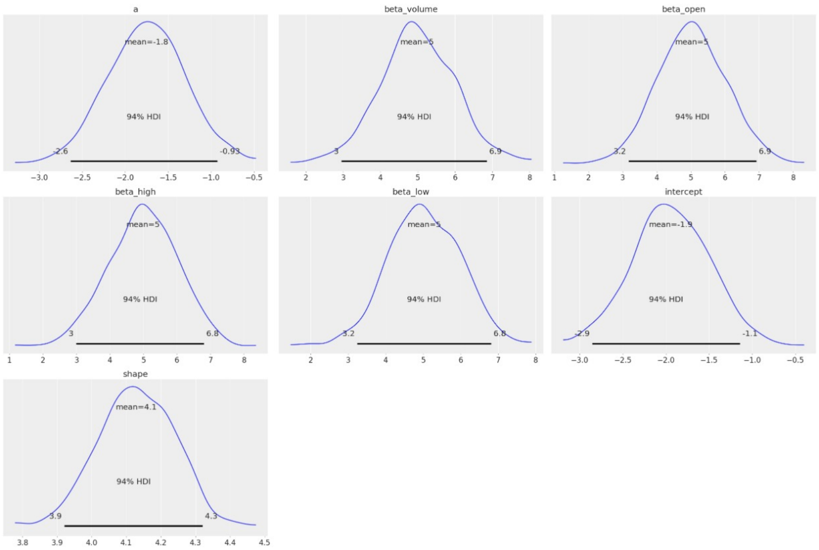

Figure 5: Density plot.