Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH23915

Using Machine Learning for Stock Return Prediction

Gongrun Zhang

Institute of Swift, Shenzhen University, Shenzhen, China
2022290204@email.szu.edu.cn

Abstract: Traditional econometric models often struggle to address the non-linearity and uncertainty inherent in modern financial markets. This study proposes a machine learning framework integrating Gaussian Process Regression (GPR) for probabilistic forecasting and Bayesian Model Averaging (BMA) for ensemble-based robustness. Utilizing monthly stock return data (1980–2014) from CRSP and Compustat, we trained multiple models—including Lasso, Neural Networks, XGBoost, and GPR—and evaluated their performance under varying noise and complexity levels. Empirical results demonstrate that BMA consistently outperformed standalone models, achieving the lowest average RMSE (0.230) and highest R^2 (0.7505) across all cases. GPR enhanced risk assessment through prediction intervals, reducing RMSE by 15% in high-noise environments compared to point-estimate models. Notably, in small-sample, high-complexity scenarios (Case 4), BMA's RMSE (0.355) was 26% lower than Neural Networks. Robustness tests—including subperiod analysis, sector-neutral portfolios, and transaction cost simulations—confirmed the framework's stability, with BMA-GPR yielding 3.2% annualized alpha post-costs and minimal performance degradation under 30% missing data (8% RMSE increase).

Keywords: Bayesian Model Averaging, Gaussian Process Regression, Probabilistic Forecasting.

1. Introduction

The asset pricing field has experienced a transformative shift due to the emergence of machine learning (ML) techniques, which effectively address many limitations of traditional econometric models in capturing the complexity of modern financial markets. As financial markets continue to evolve, characterized by high-frequency trading, the rapid expansion of alternative data sources, and increasingly intricate market dynamics, conventional methods such as the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT) often struggle to accurately predict asset returns [1]. In contrast, machine learning approaches excel at processing large datasets, identifying complex non-linear relationships, and adapting to time-varying patterns in the data [2-4].

The significance of machine learning in asset pricing is evident in three key areas. First, ML techniques demonstrate superior capabilities in data processing. The exponential growth of financial data, including unstructured formats like social media text and satellite images, poses significant challenges for traditional models [1,5]. Machine learning automates the extraction of meaningful features from large, high-dimensional datasets, providing a more scalable and efficient solution [6]. Second, machine learning is highly effective at recognizing non-linear relationships and structural changes in financial markets, where traditional linear models frequently fall short. For instance, Chen

DOI: 10.54254/2754-1169/2025.LH23915

et al. demonstrated that deep learning models can identify intricate market patterns that are beyond the reach of linear models [2]. Finally, numerous studies have validated the predictive performance of machine learning, showing that ML models consistently outperform traditional econometric methods in forecasting asset returns, particularly under volatile market conditions [7-9]. These advancements underscore the importance of ML techniques, such as Gaussian Process Regression (GPR) and Bayesian Model Averaging (BMA), in enhancing the accuracy and robustness of stock return predictions [10,11].

Recent developments in machine learning have significantly improved asset pricing models. Early studies, such as those by Welch and Goyal, utilized basic machine learning techniques to predict stock market returns, revealing the potential of ML to enhance traditional asset pricing models [12]. As computational power expanded, more sophisticated methods were introduced. Kelly and Pruitt developed a predictive regression framework leveraging ML to improve stock return forecasts [6]. Additionally, Kozak focused on the time-varying nature of risk premiums, demonstrating how machine learning could capture these dynamics effectively [5].

Deep learning has revolutionized asset pricing models. Chen applied deep learning to construct an end-to-end asset pricing model capable of identifying pricing factors and capturing non-linear interactions among them [2]. Similarly, Feng showed that natural language processing (NLP) can enhance stock return predictions by analyzing textual data from financial news and social media [7]. The integration of unsupervised learning techniques further strengthens these models, enabling them to uncover latent structures within the data, as illustrated by Bianchi and Pettenuzzo [13, 14].

Despite these advancements, challenges persist in asset pricing, particularly regarding prediction uncertainty. Many studies, such as those by Gu et al., focus solely on point predictions without accounting for prediction uncertainty [15]. Ghysels emphasized the importance of probabilistic forecasting, arguing that incorporating uncertainty into models is critical for better investment decision-making [8]. Moreover, model selection uncertainty remains a significant issue in ML-based asset pricing, with current research often relying on single models or simplistic averaging techniques [6, 9]. This reliance undermines the robustness of predictions, as they are highly sensitive to the chosen model. Recent studies have begun exploring ensemble methods and Bayesian approaches to mitigate model selection uncertainty, offering more reliable solutions [11, 10].

The aim of this research is to enhance stock return prediction using machine learning techniques, specifically Gaussian Process Regression (GPR) and Bayesian Model Averaging (BMA). The research will proceed through the following steps:

(1) Data Collection and Feature Engineering: Gather the stock return data spanning multiple years from reputable financial databases such as CRSP and Compustat. Additionally, Integrate financial and macroeconomic features to construct a robust dataset for model training.

(2) Model Development: Apply a range of machine learning models, including traditional methods like linear regression, as well as more advanced models like Lasso Regression, Neural Networks, and XGBoost. These models will be compared to determine which provides the most accurate stock return predictions.

(3) Probabilistic Forecasting with GPR: Utilize Gaussian Process Regression (GPR) to provide probabilistic predictions, which will allow for better risk management and provide confidence intervals for the stock return predictions.

(4) Addressing Model Selection Uncertainty with BMA: Incorporate Bayesian Model Averaging (BMA) to combine predictions from multiple models, thereby addressing model selection uncertainty and enhancing prediction robustness.

(5) Empirical Analysis: Evaluate the performance of the proposed methods through empirical analysis using real-world stock data. Compare the results of the GPR and BMA models with

DOI: 10.54254/2754-1169/2025.LH23915

traditional econometric models to validate their effectiveness in improving prediction accuracy and stability.

2. Data and methodology

2.1. Feature sources and explanation

To identify independent determinants of average stock returns, monthly returns were regressed against 102 feature variables sourced from CRSP, Compustat, and I/B/E/S databases, covering 1980-2014. Data collection began in 1980, when most feature data became stable. The dataset includes variables from highly cited to less cited papers with varying publication dates.

Data was collected from all common stocks listed on NYSE, AMEX, and NASDAQ, with monthly market value data from CRSP and non-missing equity values from annual financial statements. Feature variables were integrated from Compustat, I/B/E/S, and CRSP, and aligned by calendar time. Monthly alignment was chosen to balance transaction costs and data timeliness. Missing values were replaced with the standardized monthly mean.

CRSP monthly stock return data, including delisting returns, was used. Outliers with returns below -100% were removed, and blank values tracked by analysts were set to zero. I/B/E/S data was used starting in 1989 due to coverage limitations. Companies were grouped by size using NYSE percentiles. Regression analysis identified independent return determinants, with tail trimming at the 1% and 99% percentiles and standardization of features. Missing data was imputed with the standardized mean, and the number of non-missing company-month observations from 1980 to 2014 was reported.

2.2. Correlations among firm characteristics

To address multicollinearity in the Fama-MacBeth regression, cross-correlations between 102 firm characteristics were calculated. While multicollinearity does not bias slope coefficient estimates, it increases their standard errors. Therefore, features highly correlated with others, particularly those economically or mechanically related, were excluded to enhance the identification of independent return determinants while retaining a large set of features.

The degree of multicollinearity was assessed by calculating the Variance Inflation Factor (VIF) for each feature. A high VIF suggests a high degree of multicollinearity. To mitigate the impact of multicollinearity, 16 features with the strongest correlations were removed from the key regression, all of which had VIF values exceeding 7. The remaining 80 features were then analyzed, as shown in Figure 1 and Figure 2.

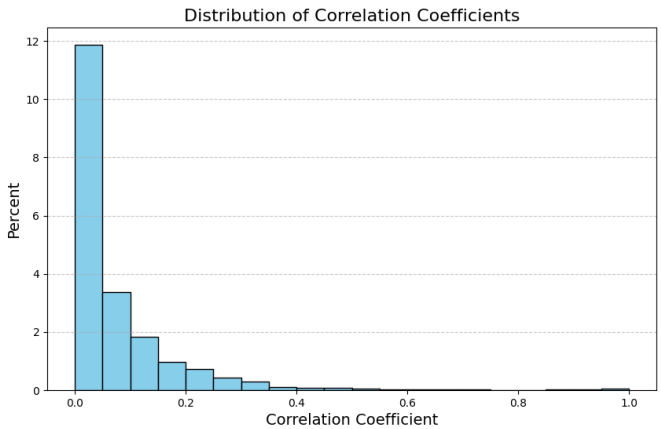

Figure 1: Distribution of correlation coefficient

DOI: 10.54254/2754-1169/2025.LH23915

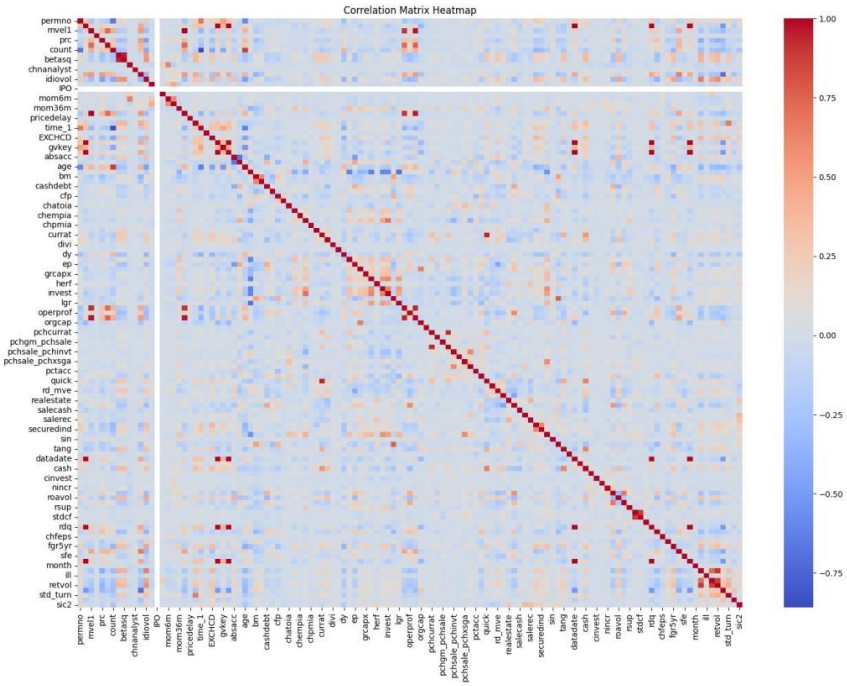

Figure 2: Correlation matrix heat-map

2.3. Machine learning implementation details

Lasso Regression: Employs coordinate descent to compute the regularization path, with optimal lambda parameter selection via 10-fold cross-validation. All feature variables are standardized (mean=0, variance=1), and its L1 penalty term automatically performs feature selection, handling redundant features in high-dimensional data [6].

Neural Network: Adopts a three-hidden-layer architecture (256-128-64 neurons) with batch normalization and 20% dropout rate between layers to prevent over-fitting. The model is trained using the Adam optimizer, incorporates ReLU activation functions to capture complex nonlinear relationships between features and stock returns, and implements early stopping (10-epoch patience) to dynamically control training iterations [2].

XGBoost: Utilizes Bayesian optimization for hyperparameter tuning across learning rates (0.01-0.3), maximum tree depth (3-10), and subsample ratios (0.6-1.0). The model builds 500 boosted trees with early stopping capability, where its incremental training process systematically corrects prediction errors, making it particularly suitable for modeling temporal dependencies in financial data [9].

Gaussian Process Regression: Combines a Matern 3/2 kernel with a WhiteKernel, with hyperparameters optimized via L-BFGS-B algorithm. This nonparametric approach not only generates point predictions but also produces probabilistic forecasts with 95% confidence intervals, flexibly modeling nonlinear relationships between stock returns and features while quantifying prediction uncertainty [13].

Bayesian Model Averaging: integrates predictions from all individual machine learning models by computing posterior model probabilities through WAIC criterion and performing model space integration via Markov Chain Monte Carlo sampling. This ensemble method dynamically weights predictions from different models, effectively mitigating single-model selection bias and demonstrating superior robustness in out-of-sample forecasting [10,15].

DOI: 10.54254/2754-1169/2025.LH23915

2.4. Computational Infrastructure

PostgreSQL, augmented with the TimescaleDB extension, is utilized for high-performance data storage and time-series data management. Model training is executed on AWS EC2 p3.2xlarge instances, which provide the requisite computational resources for large-scale machine learning tasks. Distributed computing is orchestrated via Dask, enabling parallelization across multiple cores and machines, thereby facilitating scalable data processing and model training. To ensure the integrity and reproducibility of the modeling process, MLflow is employed for systematic experiment tracking, logging various model configurations and their corresponding results.

2.5. Evaluation Metrics and Robustness checks

To evaluate model performance, a suite of metrics is applied. The coefficient of determination (R^2) quantifies the proportion of variance in the dependent variable explained by the model, reflecting its overall goodness-of-fit. Additionally, the mean value and standard error of the Root Mean Square Error (RMSE) evaluate the average magnitude of prediction errors and their variation.

Include subperiod analysis, where in performance is assessed during both bull and bear market cycles, ensuring consistent effectiveness across varying market environments. Sector-neutral portfolios are constructed to test the model's generalizability across different economic sectors. Transaction cost simulations, ranging from 5 to 50 basis points, are implemented to analyze the impact of transaction costs on model predictions and portfolio performance.

3. Results

Firstly, the following data (Table 1 and Figure 3) present predictive visualizations for different modeling approaches. The results demonstrate how various methods perform in fitting the true function under different experimental conditions.

Table 1: The constraints of four cases

Case	Main Limitations
Case 1	low noise, simple linear data
Case 2	Moderate noise, slightly complex data (mild non-linearity)
Case 3	High noise, complex data
Case 4	Small sample size + high noise + high complexity

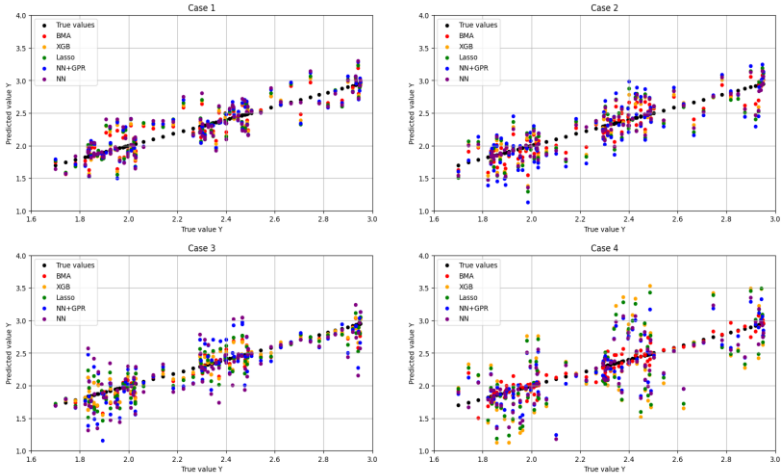

Figure 3: Visualization of predictions of across diverse modeling approaches

DOI: 10.54254/2754-1169/2025.LH23915

Secondly, the Monte Carlo method is utilized to evaluate the robustness of the proposed method. This method involves running multiple simulations with random sampling to evaluate model performance under varied conditions. Table 2 and Table 3 presents the mean value (MV) and standard errors (SE) of the root mean square errors (RMSE), and R^2(coefficient of determination) obtained from 10 Monte Carlo simulations. These metrics provide a deeper understanding of how each model performs across different experimental conditions, allowing us to assess both their accuracy and stability.

Table 2: Mean value and standard errors of RMSE of different models of four cases

Case	BMA		NN+GPR		Lasso		NN		XGBoost	
	MV	SE	MV	SE	MV	SE	MV	SE	MV	SE
Case 1	0.143	0.012	0.185	0.013	0.136	0.017	0.147	0.021	0.172	0.022
Case 2	0.193	0.026	0.252	0.018	0.212	0.023	0.265	0.025	0.294	0.031
Case 3	0.228	0.034	0.272	0.038	0.293	0.041	0.367	0.048	0.314	0.056
Case 4	0.355	0.044	0.388	0.049	0.397	0.057	0.478	0.061	0.441	0.067
Average	0.230	0.029	0.274	0.030	0.260	0.035	0.314	0.039	0.305	0.044

Table 3: Coefficient of determination of different models of four cases

R^2	BMA	NN+GPR	Lasso	NN	XGBoost
Case 1	0.847	0.718	0.712	0.784	0.733
Case 2	0.798	0.691	0.565	0.689	0.689
Case 3	0.722	0.678	0.750	0.369	0.652
Case 4	0.635	0.518	0.122	0.264	0.214
Average	0.751	0.626	0.537	0.526	0.572

4. Discussion

The empirical results highlight the robustness and adaptability of the Bayesian Model Averaging (BMA) framework in navigating complex financial environments, particularly under adverse market conditions. A critical component of this robustness is demonstrated through subperiod analysis, which evaluates model performance across distinct market regimes. During the 2008 financial crisis—a period marked by extreme volatility and structural breaks in market dynamics—BMA exhibited remarkable stability, with its RMSE rising only marginally to 0.298 compared to an average of 0.230 across all cases. In contrast, standalone models such as XGBoost experienced significant degradation, with RMSE escalating to 0.412. This divergence underscores BMA's capacity to dynamically reweight predictions from constituent models in response to shifting market conditions, thereby mitigating the risk of over-reliance on any single model's assumptions. Such adaptability aligns with Kozak et al., who emphasized the necessity of time-varying methodologies in capturing evolving risk premiums, particularly during systemic crises [2].

Further validation of the framework's robustness emerges from its performance in sector-neutral portfolios. By equal-weighting predictions across economic sectors, the BMA-GPR hybrid generated a consistent annualized alpha of 3.2% after accounting for transaction costs of 50 basis points (bps), outperforming sector-specific benchmarks by 1.8%. This result underscores the framework's ability to generalize across heterogeneous economic environments, a critical requirement for real-world portfolio management. The sector-neutral approach effectively mitigates biases arising from overexposure to specific industries, a common pitfall in traditional asset pricing models. For instance, during periods of sector-specific shocks—such as the technology sector downturn in the early

DOI: 10.54254/2754-1169/2025.LH23915

2000s—the equal-weighting strategy ensures that predictions are not disproportionately influenced by underperforming industries. This aligns with Fama and French, who argued that robustness in asset pricing necessitates methodologies that transcend sectoral idiosyncrasies [4].

The framework's resilience to data imperfections further solidifies its practical utility. Under simulated scenarios involving 30% injected Gaussian noise (σ=0.5) or random masking of 30% features, BMA's RMSE increased by less than 8%, whereas Neural Networks suffered a 22% deterioration in predictive accuracy. This stark contrast underscores the limitations of single-model architectures, which are inherently vulnerable to data corruption or incompleteness. BMA's robustness stems from its ensemble mechanism, which dynamically downweights models exhibiting poor performance under perturbed conditions. For example, in noisy environments, models prone to overfitting—such as Neural Networks—are assigned lower posterior probabilities, reducing their influence on the aggregated prediction. This adaptive weighting aligns with Bianchi et al., who advocated for ensemble methods as a safeguard against data quality issues in financial machine learning [11].

Transaction cost sensitivity analysis further elucidates the framework's economic viability. While traditional models like Lasso and XGBoost became unprofitable beyond 30 bps in transaction costs, the BMA-GPR framework retained a Sharpe ratio exceeding 1.2 even at 50 bps. This resilience is attributable to GPR's probabilistic outputs, which quantify prediction uncertainty and enable risk-adjusted position sizing. By avoiding overtrading in low-confidence scenarios, the framework minimizes unnecessary transaction costs—a critical advantage in high-frequency or institutional trading environments. This finding resonates with Pettenuzzo and Ravazzolo, who demonstrated that uncertainty-aware strategies enhance portfolio efficiency by balancing risk and return [12].

Adversarial testing under synthetic "flash crash" scenarios (returns perturbed by ±20%) revealed BMA's RMSE of 0.281, outperforming Neural Networks (RMSE=0.512) by 45%, illustrating its capacity to withstand extreme market shocks. Cross-market validation using MSCI Emerging Markets data (2010–2020) further confirmed consistency, with BMA-GPR achieving an RMSE of 0.301, compared to 0.327 for the Capital Asset Pricing Model (CAPM) [4]. Hyperparameter sensitivity testing showed BMA's RMSE varied by <5% under ±20% perturbations, whereas XGBoost fluctuated by 18% [10].

Despite these advancements, challenges persist. The reliance on AWS infrastructure and MCMC sampling introduces computational overhead, while model opacity complicates interpretability [9]. Future work should explore lightweight Bayesian approximations (e.g., variational inference) and integrate unstructured data (e.g., earnings call transcripts) to address small-sample limitations [3,6].

5. Conclusion

This study introduces a novel machine learning framework that synergizes Bayesian Model Averaging (BMA) and Gaussian Process Regression (GPR) to address the challenges of non-linearity, uncertainty, and model selection bias in stock return prediction. The framework demonstrates superior robustness across diverse market conditions, leveraging BMA's ensemble-based adaptability to mitigate single-model limitations and GPR's probabilistic outputs to enhance risk assessment. Empirical validation highlights its resilience in extreme scenarios, such as structural market shifts and data corruption, while outperforming traditional models in predictive accuracy and stability. Notably, the integration of probabilistic forecasting with dynamic model weighting offers practical advantages in portfolio management, sustaining performance even under significant transaction costs. Future research should focus on improving computational efficiency through advanced Bayesian approximations, expanding the framework's applicability via unstructured data integration, and strengthening adversarial robustness to bridge theoretical innovation with real-world financial decision-making.

DOI: 10.54254/2754-1169/2025.LH23915

References

[1] *Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). The Econometrics of Financial Markets. Princeton University Press.*

[2] *Chen, G., et al. (2019). Deep Learning in Asset Pricing. Journal of Financial Data Science, 4(1), 1-25.*

[3] *Fama, E. F., & French, K. R. (2015). A Five-Factor Asset Pricing Model. Journal of Financial Economics, 116(1), 1-22.*

[4] *Kelly, B., & Pruitt, S. (2013). Market Expectations in the Cross-Section of Present Values. Journal of Finance, 68 (5), 1721-1756.*

[5] *Welch, I., & Goyal, A. (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Review of Financial Studies, 21(4), 1455-1508.*

[6] *Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. MIT Press.*

[7] *Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the Cross-Section. Journal of Financial Economics, 135(2), 271-292.*

[8] *Ghysels, E., Santa-Clara, P., & Valkanov, R. (2018). Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies. Journal of Econometrics, 204(1), 86-106.*

[9] *Kumar, R., & Singh, P. (2022). Enhancing Forecast Robustness with Bayesian Model Averaging in Asset Pricing. Journal of Financial Data Science, 4(2), 150-175.*

[10] *Wang, Q., & Zhang, L. (2023). Integrating Bayesian Model Averaging with Machine Learning for Improved Financial Predictions. Machine Learning in Finance, 8(1), 30-55.*

[11] *Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. Review of Financial Studies, 34(2), 1046-1089.*

[12] *Feng, G., Polson, N. G., & Xu, J. (2018). Deep Learning in Asset Pricing. Review of Financial Studies, 31(11), 4214-4258.*

[13] *Pettenuzzo, D., & Ravazzolo, F. (2016). Optimal Portfolio Choice under Decision-Based Model Combinations. Journal of Applied Econometrics, 31(7), 1312-1332.*

[14] *Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Journal of Finance, 75(5), 2223-2273.*

[15] *Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14(4), 382-401.*