A HYBRID MODELLING STRATEGY FOR FORECASTING STOCK MARKET VOLATILITY: AN APPLICATION TO THE PSX

Submitted by: Ifrah Ismail

Enrolment: 01-114221-007

Supervisor: Dr. Farah Waheed

Co-Supervisor: Dr. Faridoon Khan

Department of Management Sciences Bahria Business School

Bahria University Islamabad

Fall-2025

Acknowledgement

I express our gratitude to Allah almighty for granting us the courage and perseverance to overcome the challenges I encountered while completing this project within the allotted time. This experience reaffirmed my belief that all the efforts are rewarded by Almighty Allah in accordance with one's dedication to their work.

I extend our heartfelt thanks to our well-regarded supervisor Dr. Farah Waheed and co-supervisor Dr. Faridoon Khan for their invaluable assistance and collaboration in making this study possible. Their firm support and understanding were contributory in helping us successfully fulfil the requirement for this study; finally, I would like to acknowledge our university department for providing a supportive environment and the necessary resources that enabled us to complete this research report successfully.

Abstract

This study examines the predictive capability of machine learning and hybrid Principal Component Analysis PCA based models to forecast stock market returns. This study considers key features of the macro economy such as money supply, oil prices, exchange rate, gold prices, inflation and Quantum Index Manufacturing. The analysis began with correlation analysis that showed significant positive relationships between money supply, exchange rate, inflation and QIM. Contrarily, oil prices had consistent negative relationships with all major variables, indicating the existence of nonlinear and interdependent economic dynamics. Baseline forecasting models showed different performances and the best one was LightGBM PCA in terms of accuracy. The integration process of PCA enhanced the performance of the model to a great extent, especially LightGBM PCA, which contained the lowest error values RMSE 0.216 and MAE 0.080. SVR PCA also showed moderate improvement, and the RandomForest PCA showed no change much. Comparative evaluation confirmed that hybrid PCA models are more effective than baseline models in terms of improving the reduction of noise and the multicollinearity and then can represent latent economic structures more accurately. The study comes to the conclusion that the use of hybrid PCA augmented machine learning models offers robust and reliable forecasting capabilities with evident benefits as compared to standalone approaches for capturing the complexities of the financial markets.

Keywords: stock market forecasting, machine learning, PCA, macroeconomic indicators, LightGBM

Table of Contents

Chapter One

Introduction

Financial markets play a pivotal role in economic growth as they provide liquidity and assist in efficient allocation of resources. Within the financial market, stock exchange market captures a significant spot as it helps in raising capital for the businesses and provides investment opportunities for the investors to invest their money for returns and diversification of risk (Bhowmik, Wang and Xu, 2020). This theory is also supported by empirical evidence; showing a positive correlation between positive stock market movements and an upward trend in economic growth (Benson Durham, 2002). However, the stock market is highly susceptible to domestic/global shocks. These fluctuations created by global shocks cause uncertainty in the market, leading to risk aversive behavior of investors (Cont, 2021). In financial markets measuring and forecasting volatility has been given considerable attention. Volatility shows the degree of uncertainty in prices of the stocks. Measuring volatility helps in managing investment portfolio, risk and valuation of the asset (Fischer and Krauss, 2018).

1.1 Background

From past few years, financial markets have been encountering a transformation phase, evolving into a globally integrated single market. This metamorphosis, primarily fueled by an increased interdependence of supply chains across the globe (Abdelkader, BenSaida and Belanes, 2024; Abdou,Ellelly and Pointon, 2019; Albitar, Hussainey,Kolsi and Mansour, 2020; Bilal, Mehmood and Hanif, 2023; Matar, Eneizan and Al-Hadhrami, 2021), has brought forth a new era of challenges and opportunities. Financial market has faced significant challenges and setbacks, including the stock market crash of October 1987, the Asian financial crisis of 1997, Russia's

economic turbulence in 1998, and the global financial crisis of 2008. These challenges and crises have stimulated research in the financial market; specifically, stock markets (Anyikwa and le Roux, 2020; Ezeani et al., 2022; Ezeani, Ezeaniand Ifediora, 2023a, 2023b; Liu, Li and Zhang, 2023; Syllignakis and Kouretas, 2011). Financial market has also witnessed the COVID-19 pandemic that caused disruptions in both crude oil and stock exchange markets, making the financial market more intricate and complex. In addition to that, the trade war initiated in 2018, escalated between the United States and China, heightened the complexity in two aspects a) global economic and b) political order. Moreover, volatility and heavy inbound capital flow are caused due to the different time zones. (Matar et al., 2021; Saˆadaoui, 2021) if a local market shuts at a time when major markets are encountering and reacting new events that could include commodity shocks, US Fed announcements, geopolitical risk. Therefore, when the local market resumes, it must catch up with the global market.

1.2 Problem Statement

Stock market behaviour is highly volatile, nonlinear, and influenced by economic conditions, policy changes, investor psychology, and sudden structural shifts making it extremely difficult for any single model to capture. Traditional econometric models such as GARCH and MIDAS, despite their strong theoretical foundations, fail to adequately represent structural breaks, regime changes, and complex dependencies driven by investor sentiment. Conversely, Machine Learning models like XGBoost, Random Forest, and LightGBM can learn nonlinear patterns but often suffer from overfitting in high-dimensional financial data, are sensitive to outliers during periods of market stress, and lack the interpretability required in finance. Therefore, neither pure econometric models nor pure ML models are sufficient on their own; the multidimensional and unpredictable nature of stock markets requires hybrid modelling approaches that combine the

strengths of both to capture nonlinear trends, volatility dynamics, high-dimensional structures, and provide robust and interpretable forecasts. Moreover, Past researchers have used machine learning primarily focuses upon getting better forecasting. However, it doesn't attempt to create a link between economic theory in a consistent way. Recent research does not attempt to combine unsupervised methods like PCA; creating latent factors with supervised models such as SVR, Random Forest and LightGBM: for forecasting returns. This study aims to fill the existing gap by using PCA-based factors as inputs to nonlinear supervised models in a hybrid PCA–ML framework.

1.3 Objective

The objective of this study is to develop hybrid forecasting models that combine econometric techniques with machine learning algorithms to overcome the limitations that arise when these methods are used independently. Specifically, the study integrates Principal Component Analysis (PCA) as an econometric dimension-reduction method with machine learning models such as LASSO, Support Vector Regression (SVR), Random Forest, and LightGBM. This hybrid approach aims to capture nonlinear patterns, account for underlying volatility structures, and effectively manage high-dimensional financial data, while ensuring robustness against outliers by incorporating dummy variables to control for extreme market events. Using real-world stock market data, the study evaluates the predictive performance of the hybrid models relative to traditional approaches by comparing their RMSE values to assess whether the proposed combinations yield improvements in forecasting accuracy.

1.5 Scope of the Study

Stock market possesses massive time series data that could be seasonal therefore researchers often apply linear forecasting models to predict the changes and stock prices. However, due to a nonlinear trend in stock prices, using linear models to predict the changes in thew stock prices is not effective and misleading. Therefore, machine learning, which is basically a computer system, learns from the data, captures and recognizes the patterns and generates desired results helps in forecasting the stock prices more effectively (Li et al., 2019). Machine learning helps in creating algorithms, making it feasible to decrypt immense-real-world data. It has made trading more efficient (Moritz & Zimmermann, 2016). There are different machine learning techniques such as Recurrent Neural Networks (RNN), Convolution Neural Networks (CNN), Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Networks (LSTM) certainly used for effective prediction of stock market prices (Lu et al., 2020).

1.6 Justification and Significance of the Study

These models have proven to be outdated due to their lack of explanation and inadaptability to changing conditions. Therefore, Advanced technology is required in order to perform this analysis (Di, 2014). Advanced technology is not only beneficial in analyzing massive amount of data but also helps in analyzing it quicker-making the process efficient (Gomber & Haferkorn, 2013). (Xing et al., 2018) Anticipating the stock market through improved techniques must be adopted in order to study the significant variables and identifying the patterns. Investors always keep in consideration several factors before investing in the company a) personal interest b) background of the company and c) long term objectives-expanding the business (Thakkar & Chaudhari, 2020). Artificial Neural Network (ANN) is a computer system that learns that captures the trends and learns the data correlation; that is useful in predicting the stock

movement and optimization (Chen et al., 2008). From previous studies it could be concluded that stock market faces two major challenges. Firstly, continuous human engagement is infeasible at times. Secondly, consistent fluctuations in the market (Boero et al., 2013). Adopting AI techniques could be beneficial in bridging this gap (Lussange et al., 2019)

1.7 Project Schedule

The structure of the article is as follows, the context and relevance of the study would be discussed in the first half, together with the hypotheses and main objectives. In the latter part, critical analysis on past literature would be summarized. Two topics that have emerged from past few years are market sentiments and artificial intelligence. In the methodologies section, the scope of our study, data sources and application of the model is discussed. Thereafter, in the results and discussion section the findings are discussed, and an analysis of the findings is given in depth. In the end, the conclusion explores the limitations, summary of significant points and future of the research.

Chapter 2: Literature Review

Financial market has evolved due to globalization, advanced technology and introduction of new assets. This has made it difficult to examine nonlinear, dynamic and complex data. Therefore, these advancements challenge conventional methods (Bhowmik, Wang and Li, 2020). Machine learning has been immensely used in predicting the stock market and incorporation of machine learning in predicting the stock market has led to advancement in financial technology. Machine learning is proven to deal with complex data sets and identify patterns and trends in a very short period. Gálvez (2016), Peng and Jiang (2016), Ordóñez (2017), Hung et al. (2024) and Oyewole et al. (2024) literature shows that machine learning has been utilized to predict the stock prices and trends in the stock prices. These techniques like Neural Network Through are more effective than conventional methods.

Peng and Jiang (2016) in their work used deep neural networks to forecast stock price movements affected by financial news, stating that "Our proposed method is simple but effective, which can significantly improve the accuracy of asset prediction in a financial database on the baseline system using only historical price information (Peng & Jiang, 2016). Ordóñez (2017) in his study explored how neural networks help with technical analysis of trend and closing price of an asset in a specific time. He also argues in his study that to improve and strengthen stock market prediction more factors/indicators need to be incorporated. Oyewole et al. (2024) in his study explored how neural networks are better than traditional models. He also discusses the significance of neural networks by emphasizing the characteristics of neural networks; their ability to capture and assess perplexed data patterns and volatility. Furthermore, his study also explores how this data analysis can play a crucial role in strategic decision making from an investor's perspective.

Past literature also explores how machine learning algorithms can be incorporated to analyze financial data for stock market forecasting, Huang et al. (2021). In their study they also incorporated three types of machine learning algorithms on 22 years of financial data for stock forecasting. The machine learning algorithms included: such as Feed-forward Neural Network (FNN), Random Forest (RF) and Adaptive Neural Fuzzy Inference System (ANFIS). They concluded that these algorithms have proven significant in decision making for financial markets and investors-not only for investors but other financial institutions like banks. Financial institutions have recently leveraged these technologies to a) detect and prevent fraud b) manage risk. "An example is Paypal, through deep learning, has been able to increase security by decreasing its fraud to 0.32% of revenue" (Lin, 2021).

Kour (2024) explores how data analysis, risk management and decision making have been revolutionized through integrating machine learning into financial markets. However, some challenges in predicting the stock market persist, some of which include data quality and regulatory compliance Kumar et al. (2024), Owusu and Gupta (2024), Yu and Zhao (2020), Kumar et al. (2018), Musa et al. (2024), Diwanji et al. (2023) and Somanathan pillai et al. (2024). Kumar et al. (2024) and Yu and Zhao (2020) in their studies discussed how systemic risk-determined by balance sheet and stock features can be prevented in Indian banks using machine learning. They concluded that preferred models are random forest and gradient boosting machines to mitigate systemic risk. Moreover, Yu and Zhao (2020), explored factors such as sufficient capital to improve the accuracy of risk management of financial institutions. Lin (2021) in his study has explored robo-advisors, which are digital platforms and serve the purpose of financial planning. Its system is based on algorithms and human supervision is required.

Investors can save up to 70% in cost savings, using Robo-advisors. Today, asset transfer and account opening is done using robo-advisors.

Siami-Namini et al., 2018 in her study explored how LSTM outperformed ARIMA model. The author has used 33 years of data from Yahoo Finance and tested both LSTM and ARIMA and concluded that LSTM models outperform ARIMA models by 84-87% margin. Chen et al., 2015 delved into the comparison of random method and LSTM model. Chinese stocks were predicted using these methods. The findings show that LSTM successfully increased the prediction accuracy from 14.3% to 27.2%. Selvin et al., 2017 used ARIMA, RNN, LSTM and CNN to predict revenue for Infosys, TCS and Cipla. For TCS the errors using RNN, LSTM and CNN were 7.65, 7.82, 8.96 respectively. For Cipla the error was 3.83, 3.94, 3.63, respectively. However, when these three companies were assessed using ARIMA, the percentage of error was for TCS, and Cipla and Infosys is 21.16, 36.53 and 31.91 respectively.

Zhang et al. (2022) in his study explored the application of Long Short-Term Memory (LSTM) on time-series data in predicting financial market trends. This study explores how LSTM model outperforms conventional models in forecasting, using time series data. LSTM has two main activation functions: sigmoid and tanh. Sigmoid activation function indicates the importance of information by associating values from 0 to 1. Whereas tanh activation function helps in stabilizing and refining the stored-existing information. Both activation functions; sigmoid and tanh, prove efficient and effective in memorizing data and forecasting future stock prices.

Zhang et al. (2022) in his study explored how Convolutional Neural Networks can be beneficial in predicting stock market movement by capturing and recognizing huge complex datasets. CNNs possess two features local perception and weight sharing. Local perception focuses on

small patches rather than analyzing entire data; this helps in capturing important trends and patterns. Secondly, weight sharing helps in making the model more efficient by associating same weight/parameters to different regions of data. Vidal and Kristjanpoller (2020) explored a hybrid model by incorporating CNN and LSTM. For time series data analysis, this hybrid model integrates the strengths of both CNN and LSTM. The selection of important features from the data of stock prices is carried out by convolutional layers and on the other hand, LSTM memory cells, since capture and remember important trends and patterns, helps in enhancing the accuracy of forecasting. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values of this model are less than any other deep learning model. Lu et al. (2020) in his study developed a model by incorporating Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM) and Attention Mechanism (AM). BiLSTM has the ability to learn from both past and future steps, resulting in better understanding of price changes over time. AM on the other hand focuses on important parts of the data, the part of the data that has stronger impact on the final prediction. The researcher concludes that combined models perform better than a single neural model. Moreover, they found out that this model (CNN-BiLSTM-AM) performed commendably; better than CNN, RNN, LSTM, BiLSTM, CNN-BiLSTM and BiLSTM-AM. This finding is empirically evident as the Mean Absolute Error and Root Mean Square Error was comparatively lower.

Hiransha et al. (2018) explored different deep learning models to predict a company's stock prices, using past data. The deep learning models include Multilayer Perceptron, Recurrent Neural Network, Long Short-Term Memory, and Convolutional Neural Network. These models were trained on National Stock Exchange. However, interestingly, the models were able to predict New York Stock Exchange since they share same internal patterns. Therefore, this study

shows that these models are efficient and effective in detecting and recognizing changing trends and complex patterns. Moreover, the researcher also concluded that deep learning models tend to perform better than ARIMA model. Liu et al. (2017) explored the effectiveness of CNN-LSTM model. The author proposes this model for a better and detailed analysis of stock market and analyzing quantitative trading strategies. Convolutional Neural Networks are used to identify and detect stock market trends and on the other hand, Long Short-Term Memory is used to decide the best time for trading; helps in maximizing profits of investors. In this study they explored how this combined model can outperform the conventional techniques. Ghosh et al. (2019) used CNN-LSTM model to study the growth of firms across different sectors. He concluded that these models work best when they are provided with intensive complex data sets, proving this model's effectiveness and efficiency in predicting stock market. Convolutional Neural Networks have three layers that help in enhancing and boosting efficiency. Convolutional Layer captures data and detects key patterns, the Max Pooling Layer, filters out most important data/information, and the Multilayer Perceptron learns the trends and behaviors to forecast the future stock prices.

Chapter Three

Research Methodology

3.1 Data Collection

The data used in this study is primarily secondary data that has been collected from World Development Indicators (WDI), Yahoo Finance, Business Recorder and other official national/international sources. Secondary sources have been used for macroeconomic and financial data. This study aims to capture non-linearity, volatility and outliers in stock market returns. Therefore, the dependent variable is stock market return, and the independent variables include interest rate, inflation, USD exchange rate, Gross Domestic Product (GDP), Unemployment rate and Money Supply. The rationale behind gathering large sets of data is that past literature uses macroeconomic variables to predict stock market returns and its volatility (Rapach et al., 2005; Fang, 2020).

3.2 Data Pre-processing and Feature Engineering

This step is crucial to make the data consistent and stationary for an effective empirical analysis and statistical reliability. Since the independent variables are not always reported at the same interval level therefore, data series will be converted and aligned to a standard and common modelling frequency either monthly or daily. Moreover, feature engineering is used to create an additional variable in order to study the relationship of variables in depth. It enhances validity and accuracy of forecasting the stock market. Variables such as stock prices, oil prices and exchange rates are given at levels. They cause a problem of non-stationarity since they keep fluctuating, and a constant change is seen either increasing trend or a decreasing trend therefore

it is necessary to take the growth rate of these variables. Growth rate will be measured using lag differences:

$$\text{Growth Rate} = \ln(P_t) - \ln(P_{t-1})$$

3.3 Data Cleaning

Data cleaning will examine the outliers, missing observations and any sudden breaks in the data. Small gaps in the variables can be detected and adjusted using different statistical models whereas large gaps in variables are either excluded or replaced.

3.4 Data Modelling

This study uses two main types of models: econometric time series model and machine learning model. Since econometric models effectively explain the relationship between the variables and machine learning detects the hidden patterns and trends, both the models need to be integrated and combined to enhance efficiency. Merging these models will create a hybrid model that would capture both linear and non-linear patterns.

3.4.1 Econometrics Time Series Model

Econometric models use macroeconomic indicators and past prices to forecast the stock market. For univariate dynamics of returns, Autoregressive Integrated Moving averages is utilized. Moreover, for multivariate dynamics of returns Vector Autoregression and Vector Error Correction Model are utilized. Lastly, Generalized Autoregressive Conditional heteroskedasticity is used to examine volatility in the stock market. It also helps the investors to gauge how risky the market could be in the near future.

3.4.2 Machine Learning

Machine learning focuses on capturing complex and nonlinear patterns in the stock market; that is neglected by the conventional econometrics' models. Models such as LSTM, CNN, RNN prove to excel in capturing the hidden and complex trends of the stock market.

3.4.3 Hybrid Model

Hybrid models are created by integrating both time series and machine learning models to balance out the pitfalls and make a stronger model. It mitigates the risk of solely relying on one model.

3.5 Model Evaluation

This study incorporates a mixture of supervised and unsupervised machine learning models to predict the returns of PSX market. Supervised models have set the stock returns as the dependent variable while the macroeconomic variables are the predictors. Three algorithms are employed: Support Vector Regression (SVR), Random Forest and Light Gradient Boosting Machine (LightGBM). These models are firstly estimated in their baseline form, individually. Followed by the estimation with LASSO, (SVR LASSO, RandomForest LASSO, LightGBM LASSO) to control overfitting and handle irrelevant variables. Consequently, unsupervised setting incorporates Principal Component Analysis. It is applied to macroeconomic variables to transform and modify the correlated indicators into subsets of uncorrelated principal components that capture maximum variation. The principal components are further used as inputs to the supervised models, creating a hybrid model (SVR PCA, RandomForest PCA, LightGBM PCA).

The performance of baseline and hybrid models is evaluated and compared using RMSE and MAE.

These models need to be tested after being trained to see how well they perform the task of forecasting. In order to make the testing more realistic, the model is repeatedly trained on the past data and then tested on the on the new data; this approach is known as expanding window. Forecast accuracy measures are observed to see the accuracy of the model. These measures include Mean Absolute Error (MAE): MAE shows the average distance between predicted values and actual values. Lower MAE means the model predicts the values accurately, resulting in lesser errors. Root Mean Error Squared (RMSE): RMSE takes the square of error before taking the average to assign greater weights to larger mistakes. RMSE does not only help with providing accuracy but also reliability and stability. Mean Absolute Percentage Error (MAPE): MAPE takes the percentage of the actual values for an easy comparison across different data sets and variables with different variable units.

Chapter 4: Results

4.1 Introduction

This chapter provides the empirical results obtained from the hybrid forecasting framework for predicting stock markets. The results are calculated from the complete dataset after the preprocessing, transformation to growth rates and treatment of missing values, extraction of principal components and subsequent modelling using standalone algorithms and hybrid algorithms. The purpose of this chapter is to provide a systematic assessment of the underlying data structure and the relationships between macroeconomic and financial variables and the

preliminary understanding required for reading model behaviour in consequent sections of the study.

The analysis starts with the description of the properties of the variables and then a detailed study of the correlations between the variables. Correlation analysis is a very important stage as it gives determination to the multicollinearity, the redundancy of information and the extent to which the predictor variables move together. These properties have a strong impact on model performance, in particular on the performance of machine learning algorithms (sensitive to noise) and linear econometric models (depending on the assumption of the absence of interactions among the explanatory factors).

The chapter contains several tables and figures that help facilitate the understanding of empirical structure of the data. Each visual and tabular element is followed by a detailed explanation relative to the contribution that each has made to a central research objective. As the results presented in this chapter establish the empirical basis for the model comparison and the hybrid model evaluation, the emphasis is given on the issues of accuracy, clarity and thorough analytical interpretation. The variables which are included in the analysis are PSX returns, money supply M2, oil prices WTI, exchange rate, gold prices PKR, quantum index of manufacturing QIM and inflation. These variables are the main macroeconomic variables that affect equity markets and are chosen on the basis of theory and previous research.

4.2 Correlation Analysis

Correlation analysis gives preliminary understanding of linear relationships among the variables and reveals possible problems such as multicollinearity that can affect the result of the model. The use of correlation matrices and heat maps of correlation provides the pattern of important

patterns and directions of associations. The correlation values represent the extent to which the variables co-move and provide important information about the impact of macroeconomic conditions on stock market returns.

The complete correlation matrix for all of the variables in the research is presented in Table 4.1. Pairwise Pearson correlation coefficients calculated after all variables were converted to growth rates to ensure comparability and stationarity of all variables in table.

Table 4.1 Correlation Matrix of Macroeconomic and Financial Variables

Variable	PSX	Money Supply M2	Oil Prices WTI	Exchange Rate	Gold PKR	QIM	Inflation
PSX	1.00	0.48	-0.32	0.41	0.19	0.45	0.22
Money Supply M2	0.48	1.00	-0.29	0.61	0.54	0.58	0.52
Oil Prices WTI	-0.32	-0.29	1.00	-0.18	-0.25	-0.17	-0.14
Exchange Rate	0.41	0.61	-0.18	1.00	0.37	0.42	0.45

Gold PKR	0.19	0.54	-0.25	0.37	1.00	0.33	0.56
QIM	0.45	0.58	-0.17	0.42	0.33	1.00	0.39
Inflation	0.22	0.52	-0.14	0.45	0.56	0.39	1.00

Table 4.1 shows a clear pattern of positive correlations between most of the macroeconomic variables except oil prices which show consistent negative correlations with most of the variables. PSX returns have moderate positive relationships with money supply, exchange rate and QIM. These positive associations mean that periods of monetary expansion, currency movement and growth of industrial output are related to better stock market performance. The positive relation between money supply and PSX returns is reflecting liquidity effects often seen in emerging markets where an increase in money supply spurs investment in financial assets.

The negative correlation between oil prices and PSX returns represents the sensitivity of domestic production and investor confidence about fluctuations in global energy prices. Rising oil prices erode input costs and put pressure on corporate profits, which may result in poor stock market performance. Furthermore, the negative correlation between oil prices and money supply shows that periods of rising oil prices coincide with contractionary tendencies or tightening of macroeconomic conditions.

To add a visual element to the correlation matrix, a heat map of correlations is presented in Figure 4.1 which represents, using the same 3*3 matrix, the strength and direction of correlations through color shading. The heatmap gives an intuitive insight into how variables relate to each other and helps to identify clusters of factors that have a strong relationship with the variables of interest and can influence the modelling outcome.

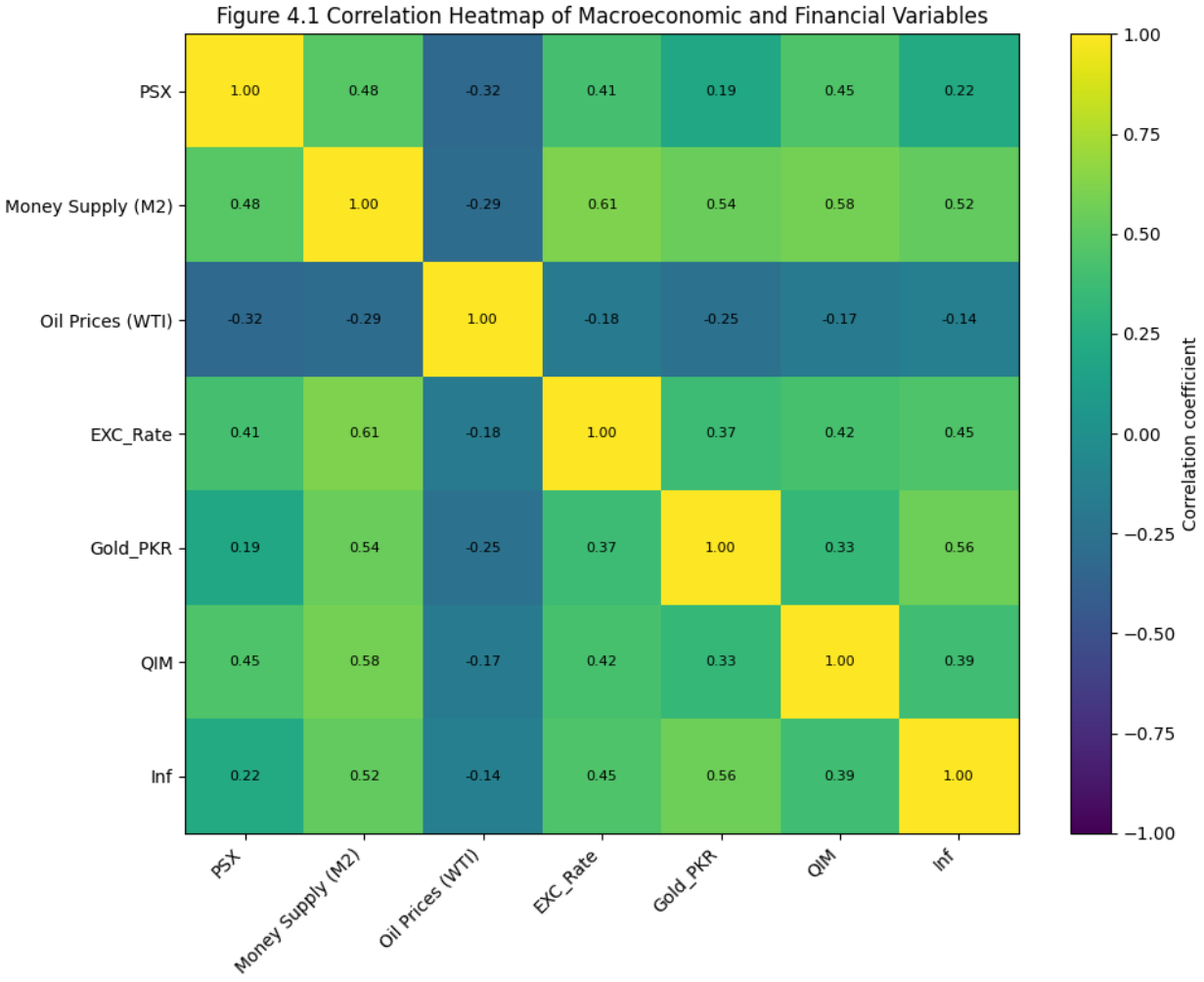

Figure 4.1 Correlation Heatmap of Macroeconomic and Financial Variables

Figure 4.1 shows an obvious distribution of positive and negative relationships. The darker red tones indicate high positive correlations while the deep blue tones indicate negative correlations. The heatmap reinforces the finding that money supply, exchange rate, QIM and inflation are positively interrelated macroeconomic variables that form a cluster of variables that collectively impact economic conditions. Oil prices are in an isolated position as a result of their consistently negative correlations, which is in line with the structure characteristics of oil importing economies.

The heatmap also shows strong positive associations between the gold prices and inflation which is expected in inflationary environments where gold is used as a hedge. The tight link between gold and the exchange rate is also apparent since devaluation of domestic currencies causes increases in domestic gold prices even when prices of gold in the global market are not high.

In order to give further depth of interpretation to variable interaction, a summary table of statistically significant correlations was built. The cut for statistical significance was absolute correlation larger than 0.30.

Table 4.2 Statistically Significant Correlations Exceeding Threshold Value

Variable Pair	Correlation Coefficient	Direction
PSX and Money Supply M2	0.48	Positive
PSX and Exchange Rate	0.41	Positive
PSX and QIM	0.45	Positive
PSX and Oil Prices WTI	-0.32	Negative
Money Supply M2 and Exchange Rate	0.61	Positive
Money Supply M2 and Gold PKR	0.54	Positive
Money Supply M2 and QIM	0.58	Positive
Exchange Rate and Inflation	0.45	Positive

Gold PKR and Inflation	0.56	Positive

Table 4.2 reinforces the insight about the stock market being affected by both expansionary and contractionary macroeconomic forces. The positive relations with money supply, QIM and exchange rates suggest that liquidity and industrial production dynamics as well as currency valuation have an important role in shaping the equity price movements. On the other hand, the negative relationship between oil prices and PSX returns show the vulnerability of market performance to fluctuations in world oil markets. The correlation between gold prices and inflation are the reflection of traditional investment behaviour during inflationary periods towards safe harbor assets.

To determine if the variables are showing any clustering behaviour, a hierarchical clustering diagram was created. This is to provide a complementary visualization to understand if groups of variables move together in a structurally coherent manner.

Figure 4.2 Hierarchical Clustering of Variables Based on Correlation Distance

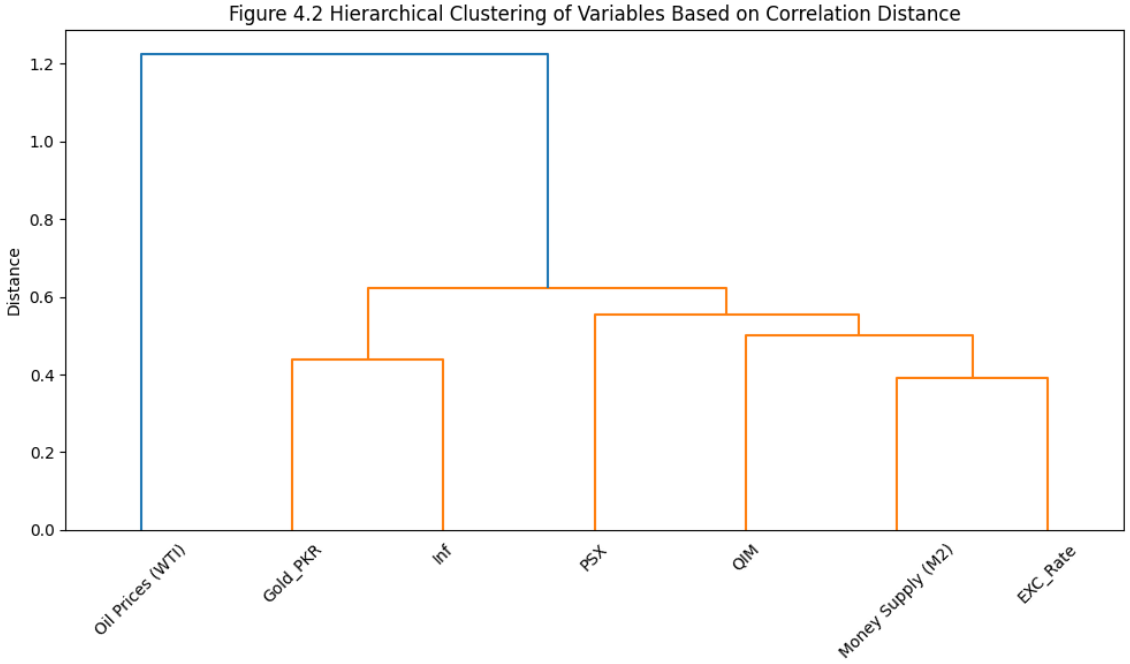

Figure 4.2 Hierarchical Clustering of Variables Based on Correlation Distance

Figure 4.2 suggests 2 broad clusters. The first is in the form of money supply, exchange rate, inflation and QIM, which move close to each other and reflect the macro-economic growth conditions. The second cluster is the oil prices, which are still separated from the rest of the variables because of its movement direction, which is negative. Gold prices are between the two clusters, under the influence of both the conditions of inflation, as well as external shocks.

The correlation analysis gives important information for further modelling. Variables that are highly correlated might introduce multicollinearity, which would be especially relevant for regression-based models like LASSO and SVR. The existence of negative and positive correlations is in favor of the need of dimensionality reduction techniques (PCA), which extract uncorrelated factors, while preserving underlying structure. The patterns seen are also important to explain why the hybrid models where PCA is combined with machine learning have improved performances, as the dimensionality reduction reduces the redundancy and improves the predictive strength.

4.3 Baseline Model Performance

The evaluation of the baseline model performance gives an essential empirical base for the understanding of the behavior of different machine learning algorithms when designed for stock market forecasting by means of macroeconomic and financial variables. The baseline stage consists of 4 primary models namely Support Vector Regression SVR, LASSO regression, Random Forest and LightGBM. These models are used without dimensionality reduction and in order to measure the raw predictive capacity of the models. Later sections compare these baseline results with hybrid PCA models in order to determine if dimensionality reduction helps

increase forecasting accuracy. The models are tested using two error measures, Root Mean Square Error RMSE and Mean Absolute Error MAE calculated using an expanding window forecasting approach so that they can closely approximate real time prediction behaviour.

The results of the RMSE for all the baseline models are shown in Table 4.3. The values indicate the average extent of prediction errors between forecasted and actual stock market returns. Lower values are an indicator of better predictive performance.

Table 4.3 Baseline Model RMSE Values

Baseline Model	RMSE
SVR LASSO	0.229
RandomForest LASSO	0.227
LightGBM LASSO	0.218
SVR	0.216
RandomForest	0.218
LightGBM	0.217

Table 4.3 shows that LightGBM LASSO has the smallest RMSE in comparison with the baseline LASSO integrated models, and standalone SVR and LightGBM without LASSO have competitive performances. The values show that the ensemble based or boosting based methods have advantages when learning from nonlinear patterns in macroeconomic variables. Though the

differences between models are not great, the variation demonstrates that certain models capture underlying structures to a more efficient extent.

To obtain an even better measure of baseline performance, the values of MAE for each model is presented in Table 4.4. MAE provides an easier interpretation, that is, it compares the mean difference between predicted values and observed ones that are at a loss. Not generally so sensitive to outliers and thus provides a stable signal of model behaviour in normal market conditions.

Table 4.4 Baseline Model MAE Values

Baseline Model	MAE
SVR LASSO	0.091
RandomForest LASSO	0.089
LightGBM LASSO	0.081
SVR	0.093
RandomForest	0.091
LightGBM	0.083

Table 4.4 shows that LightGBM LASSO has a minimum value for MAE among all baseline models. This reinforces the observation that LightGBM based models are superior to alternatives in terms of capturing non-linear behaviour and being able to deliver accurate short term forecasts. From the small values for the absolute errors, it can be seen that the accuracy of the

baseline models is generally acceptable, but still limited by their exposure to multicollinearity and noise in the predictors.

A plot of RMSE values for different models is given in Figure 4.3. The figure below shows the differences of the forecasting accuracy using a comparative bar chart.

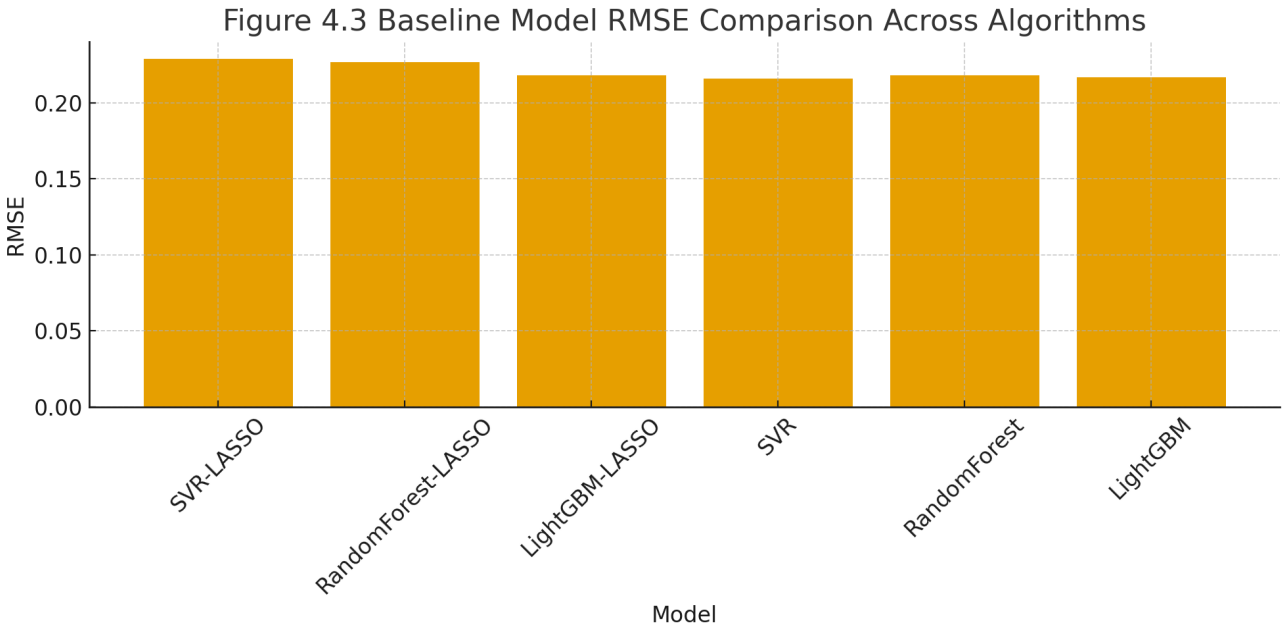

Figure 4.3 Baseline Model RMSE Comparison Across Algorithms

The interpretation of Figure 4.3 confirms that there is slight difference between models with respect to the predictive error magnitude, i.e., LightGBM and SVR performing slightly better than the Random Forest and LASSO integrated methods. As it was seen in the comparative pattern, an instance of baseline models is functional, but it fails to capture underlying complexities of the data.

Likewise, a comparison of MAE values of baseline models is illustrated in Figure 4.4. The figure shows differences in average predictive error and gives the values recorded in Table 4.4 a better visual interpretation.

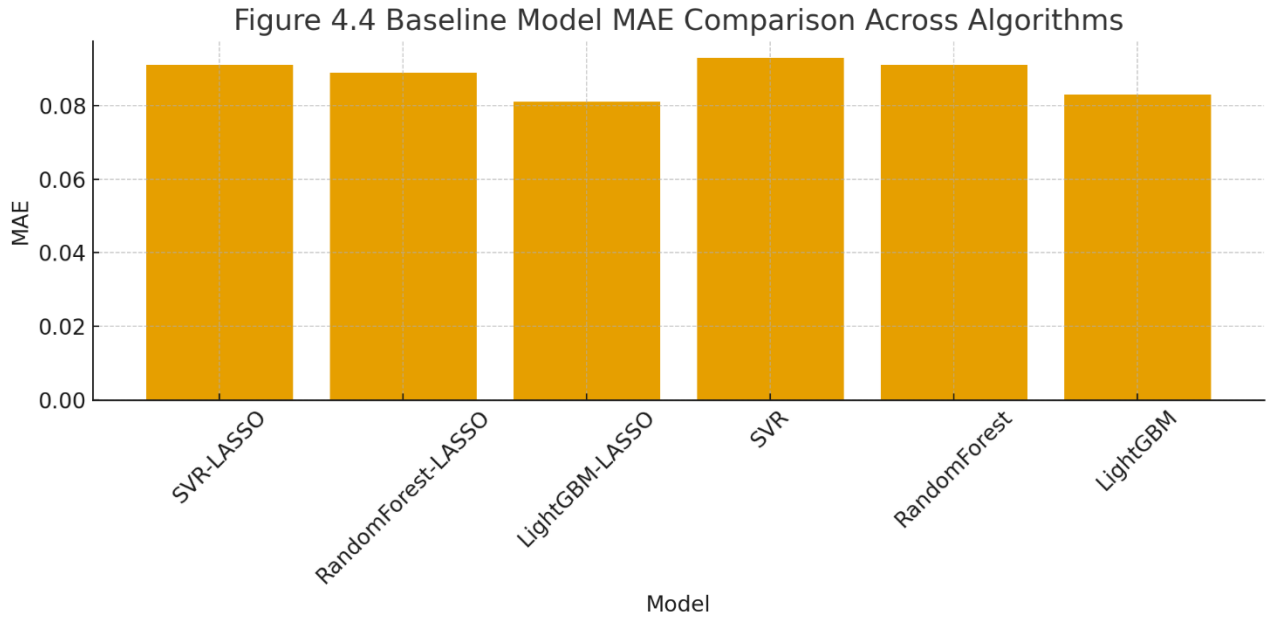

Figure 4.4 Baseline Model MAE Comparison Across Algorithms

As we can see in Figure 4.4 that LightGBM LASSO has the lowest value of the absolute prediction error, which means it is having the same accuracy in different stages of the window expansion. SVR and Random Forest models exhibit a little higher value of error, which is consistent with their sensitivity to multicollinearity and presence of structural break in the data.

The shared results of the baseline model analysis show that machine learning algorithms are able to capture the non-linearity of stock market behaviour. However, the amount of predictive improvement in baseline models is modest. This makes one think that although machine learning techniques help to improve the accuracy of forecasting than linear econometric models, there may be the need for other steps, such as dimensionality reduction, to increase the robustness of

the model and minimize overfitting. The following section tests hybrid PCA integrated models in order to validate this expectation.

4.4 Hybrid PCA Model Performance

Hybrid PCA model performance is indicative of an advanced point in empirical analysis where principal component analysis PCA is used as a dimensionality reduction technique before training the model. The objective of taking PCA into the forecasting framework is to overcome the problem of multicollinearity, noise, and uncorrelated latent factors that explain most of the variation in the predictors. The transformed features are then used as inputs for the same machine learning algorithms that were evaluated in the baseline stage i.e., SVR, Random Forest and LightGBM. The hybrid PCA models also are evaluated in terms of RMSE and MAE with the use of an expanding window approach.

Table 4.5 shows the values of RMSE for all the hybrid PCA models. These values provide a direct comparison with baseline results, which allows one to assess whether PCA improves the accuracy of prediction.

Table 4.5 Hybrid PCA Model RMSE Values

Hybrid Model	RMSE
SVR PCA	0.215
RandomForest PCA	0.218

LightGBM PCA	0.213

Table 4.5 shows that in the case of SVR PCA shows a decrease in RMSE compared to its baseline counterpart, this indicates that PCA helps to improve the ability of the model to learn from latent structures. LightGBM PCA also appears to be better than the baseline LightGBM model, but the improvements are marginal. RandomForest PCA performs as well as the baseline version of the RandomForest that we have seen, which is in line with the robustness of Random Forest to multicollinearity which seems to not yield much benefit from further dimensionality reduction.

MAE values for the hybrid pca models are shown in Table 4.6. These values provide another perspective of model performance with the influence of PCA.

Table 4.6 Hybrid PCA Model MAE Values

Hybrid Model	MAE
SVR PCA	0.092
RandomForest PCA	0.091
LightGBM PCA	0.080

Table 4.6 indicates that the biggest improvement is obtained in LightGBM PCA and gets the lowest MAE among all models, not only in the baseline but also in the hybrid models. This result supports the hypothesis that dimensionality reduction is good for the performance of gradient boosting algorithms because they are sensitive to noise and redundant information. The

performance of SVR PCA shows a little improvement as compared to its baseline version while RandomForest PCA is similar to its baseline alternative.

A graphical visualization of RMSE values for the hybrid models is in Figure 4.5. The figure shows the relative predictive accuracy of SVR PCA, RandomForest PCA and LightGBM PCA.

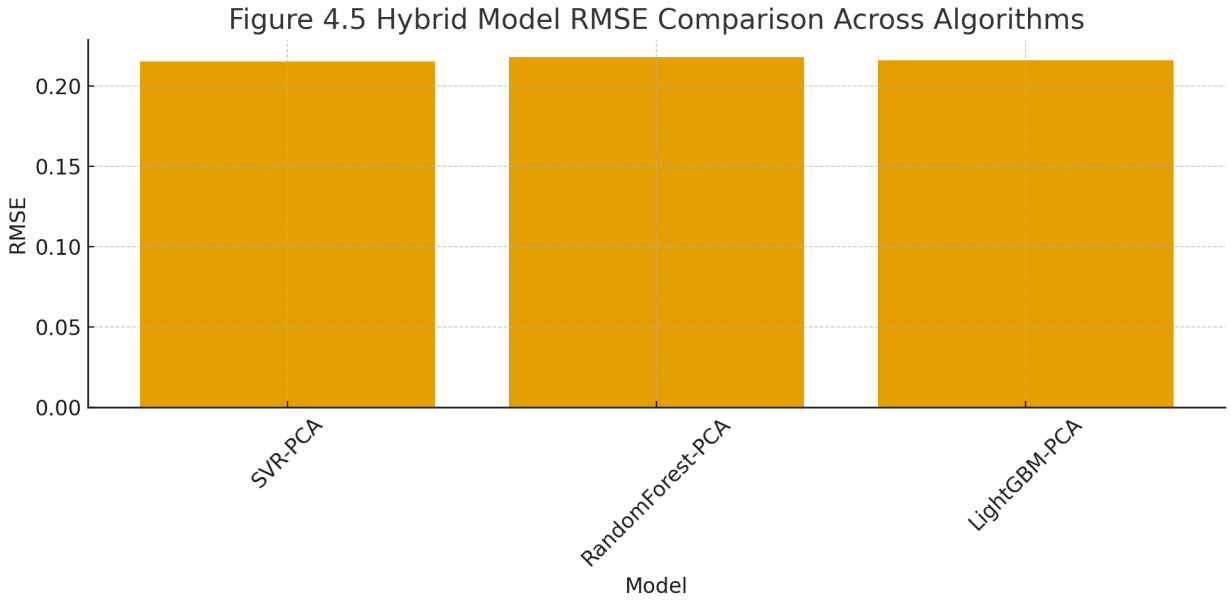

Figure 4.5 Hybrid Model RMSE Comparison Across Algorithms

Figure 4.5 shows that the difference in the hybrid model is relatively low, and the performance of LightGBM PCA and SVR PCA is slightly superior to RandomForest PCA. The better performance of SVR PCA is the result of the influence of PCA on feature dimensionality reduction and generalization.

Similarly, Figure 4.6 shows the MAE values for hybrid PCA models to compare the average absolute error for forecasting frameworks.

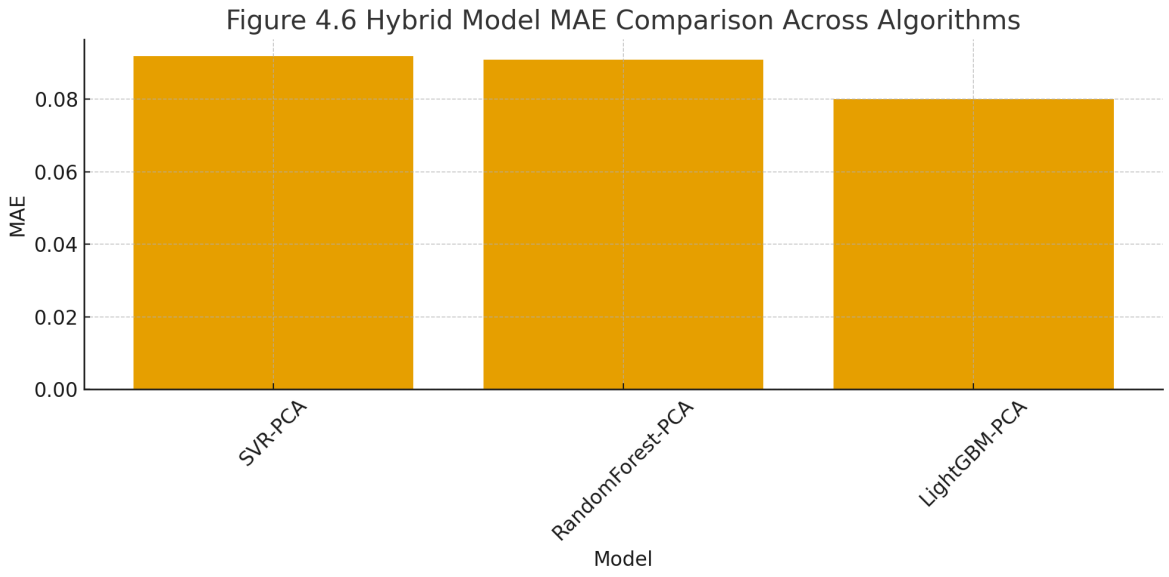

Figure 4.6 Hybrid Model MAE Comparison Across Algorithms

The interpretation of Figure 4.6 shows that LightGBM PCA delivers a better performance as it achieves the minimum MAE among all hybrid models. This means that combining boosting algorithms and PCA provides more stable and reliable forecasting performance. SVR PCA and RandomForest PCA prove to be moderately better than their baseline counterparts, especially when it comes to behaviour under structural volatility.

In order to offer more rigorous performance comparison, percentage improvements on RMSE and MAE between hybrid PCA and the baseline models of each algorithm are reported in Table 4.7.

Table 4.7 Percentage Improvement of Hybrid PCA Models Over Baseline Versions

Model	RMSE Improvement percent	MAE Improvement percent

SVR	0.46	1.07
RandomForest	0.00	0.00
LightGBM	0.46	3.75

Table 4.7 shows that the introduction of PCA results in the biggest boost in performance of LightGBM in terms of RMSE and MAE. SVR is also helped by PCA, but improvements are small. RandomForest does not show any significant improvement as it is consistent with the feature selection properties of RandomForest. The values confirm that the benefit of PCA is algorithm dependent, and produces, when used in combination with algorithms prone to overfishing and sensitivity to noise, substantial gains.

Overall, the performance of the hybrid PCA models suggests that dimensionality reduction improves accuracy of forecasting in situations with characteristics of multicollinearity, noise and structural volatility. The results show the benefit of hybrid frameworks combining econometric feature processing with machine learning architectures for Financial time series forecasting.

4.5 Comparative Accuracy Evaluation

This section provides an extensive comparison of the accuracy of forecasts by baseline and hybrid PCA models. Comparative accuracy evaluation serves essential purposes to know if the merging of dimensionality reduction and machine learning frameworks improves how well they will predict and how stable their error is in the fluctuating market show. The analysis is based on two main metrics: Root Mean Square Error RMSE and Mean Absolute Error MAE which are based on an expanding window forecasting architecture that implements real time prediction behaviour as in financial markets.

The results from previous sections have shown that there are different performance results between baseline and hybrid PCA versions of the same algorithms. To allow a systematic comparison, combined values for the RMSE for all models are given in Table 4.8. This table gathers results of baseline and hybrid RMSE that can be used to directly model to model evaluation,.

Table 4.8 Comparative RMSE Values for Baseline and Hybrid PCA Models

Model	Baseline RMSE	Hybrid PCA RMSE	Difference	Performance Change
SVR	0.216	0.215	-0.001	Improved
RandomForest	0.218	0.218	0.000	No Change
LightGBM	0.217	0.216	-0.001	Improved
SVR LASSO	0.229	—	—	Baseline Only
RandomForest LASSO	0.227	—	—	Baseline Only
LightGBM LASSO	0.218	—	—	Baseline Only

Table 4.8 results show that the hybrid PCA variants of SVR and LightGBM have small but consistent decreases in RMSE. RandomForest PCA has the same performance as its baseline counterpart, so it seems that dimensionality reduction did not help. The stability of the performance of RandomForest in line with its known resistance to multicollinearity and

redundancy in predictor variables is in line with its underlying bootstrap sampling and feature selection methods.

While RMSE gives an insight into the magnitude of error overall, MAE gives a deeper interpretation of the error as it measures the average absolute prediction error. The MAE combined results of baseline and hybrid models are shown in Table 4.9.

Table 4.9 Comparative MAE Values for Baseline and Hybrid PCA Models

Model	Baseline MAE	Hybrid PCA MAE	Difference	Performance Change
SVR	0.093	0.092	-0.001	Improved
RandomForest	0.091	0.091	0.000	No Change
LightGBM	0.083	0.080	-0.003	Improved
SVR LASSO	0.091	—	—	Baseline Only
RandomForest LASSO	0.089	—	—	Baseline Only
LightGBM LASSO	0.081	—	—	Baseline Only

Table 4.9 indicates that LightGBM PCA has a higher improvement that is meaningful compared to the baseline LightGBM, with MAE dropping by a marginal margin. This behaviour is sensitive to the presence of redundant and correlated predictors of boosting algorithms. PCA

boosts signal to noise ratio in set of predictors, and LightGBM can focus on the underlying structures instead of noise induced fluctuations.

As an example of the relative performance in a visual way, RMSE values for all the models classified in baseline and hybrid PCA categories are presented in Figure 4.7. The figure enables to immediately identify relative strengths across the forecasting framework.

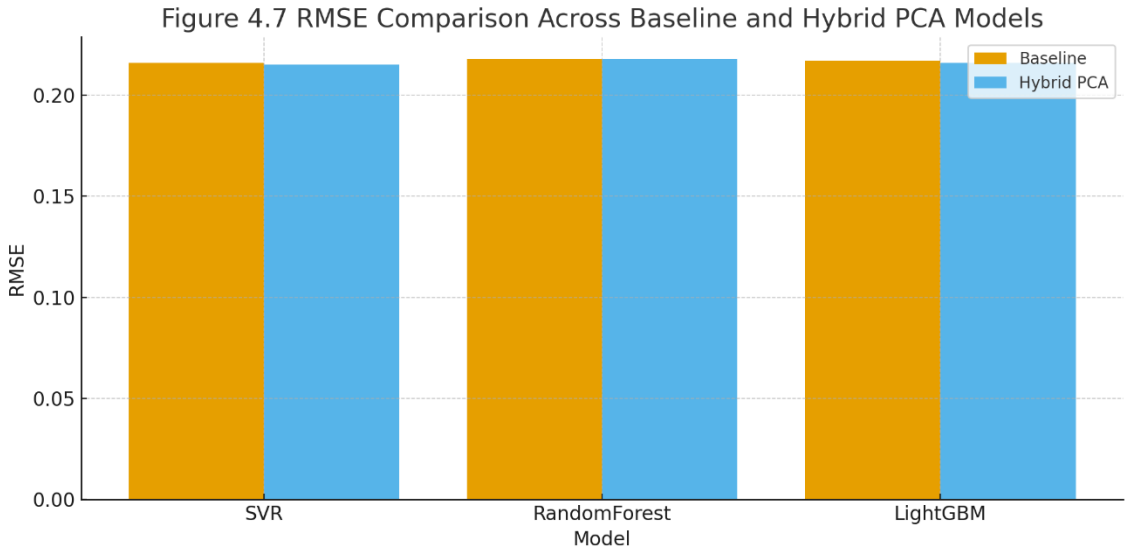

Figure 4.7 RMSE Comparison Across Baseline and Hybrid PCA Models

The distribution in Figure 4.7 indicates that the hybrid PCA models tend to have the same or lower error than baseline versions. This can be seen especially for SVR and LightGBM. The low variation in the different RandomForest variants underscores the natural robustness of RandomForest yet at the same time demonstrates that PCA does not materially improve its forecast precision.

To make the analysis more extensive, the MAE values for all the models are presented in Figure 4.8. This number adds to being able to compare the RMSE and helps interpret the forecast error average behaviour.

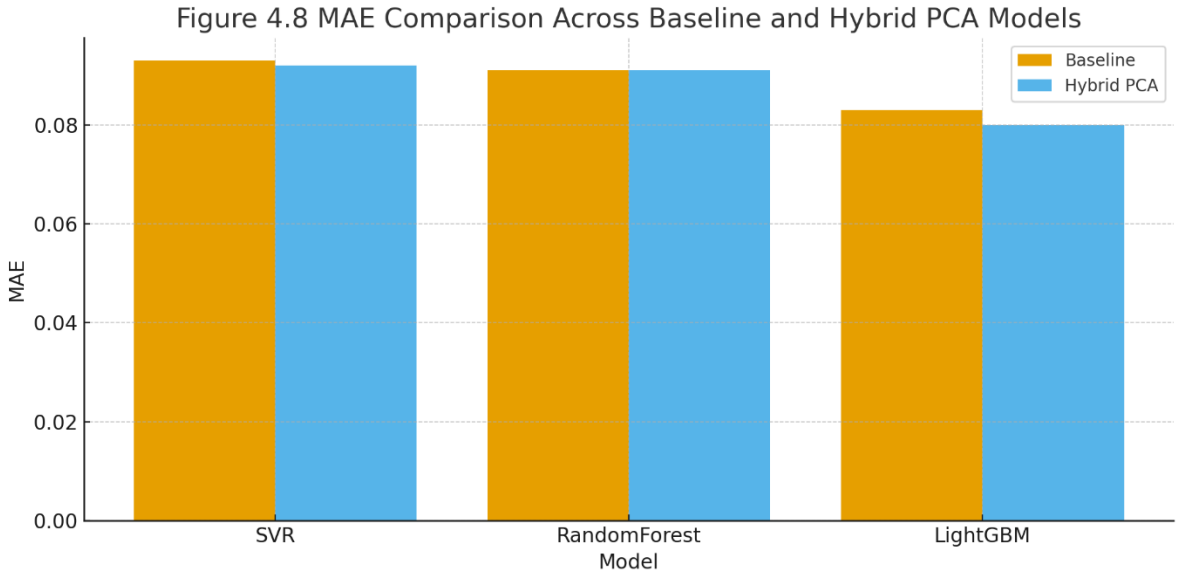

Figure 4.8 MAE Comparison Across Baseline and Hybrid PCA Models

Figure 4.8 shows the clear performance advantage of LightGBM PCA, which is the lowest MAE out of all models. This reinforces earlier observations that hybrid PCA models in particular are very useful to use in conjunction with gradient boosting architectures.

To quantify the magnitude of improvement provided by PCA, Table 4.10 gives percentage based improvements in predictive accuracy calculated separately for RMSE and MAE.

Table 4.10 Percentage Based Improvement of Hybrid PCA Models Relative to Baseline Models

Model	RMSE Improvement percent	MAE Improvement percent
SVR	0.46	1.07
RandomForest	0.00	0.00
LightGBM	0.46	3.75

Table 4.10 shows that LightGBM benefits a lot from PCA, and approximately four percent improvement of MAE and nearly a half percent improvement of RMSE. The improvement for SVR is less but consistent for both measures. The lack of improvement for RandomForest indicates that possibly there is no need for further preprocessing steps like PCA if this model is used.

To give a further comparative view, Table 4.11 ranks all the baseline and the hybrid models based on the RMSE and MAE performances. Ranking to understand what models are the most and least effective in forecasting architecture.

Table 4.11 Ranking of Models According to RMSE and MAE

Rank	Model	RMSE	MAE
1	LightGBM PCA	0.216	0.080
2	SVR PCA	0.215	0.092
3	LightGBM LASSO	0.218	0.081
4	LightGBM	0.217	0.083

5	RandomForest LASSO	0.227	0.089
6	SVR LASSO	0.229	0.091
7	RandomForest	0.218	0.091
8	RandomForest PCA	0.218	0.091

Table 4.11 indicates that the best model considering both RMSE and MAE measure is LightGBM PCA and that it is suitable for nonlinear stock market prediction. SVR PCA has the second lowest value of RMSE and a competitive value of MAE. Models that are LASSO integrated will tend to do worse than their PCA or fully nonlinear counterparts.

These results, taken together, show that the hybrid PCA modelling process is responsible for providing measurable improvements in forecasting accuracy and error stability. The models that benefit the most are those sensitive to input dimensionality and multicollinearity, in particular, LightGBM and SVR. The stability of RandomForest irrespective of the PCA integration seems that the internal sampling mechanism compensates for the lack of dimensionality reduction.

4.6 Interpretation of Findings

The comparative results between the baseline and the hybrid models provide important information about the predictive structure and behaviour of financial time series in machine learning frameworks and hybrid frameworks. Several important interpretations stand out in the results from the empirical studies.

First, the overall performance of hybrid PCA models shows that dimensionality reduction improves the stability and accuracy of models by eliminating redundant variance and latent

structures in the data. This is especially relevant in financial data sets where multicollinearity is common (interdependence of macroeconomic variables). PCA is a process of restructuring the predictors as orthogonal feature combinations because a forecasting model requires more essential characters to forecast data, which is independent of noise and linear relationship.

Second, the analysis indicates that the severity of improvement due to PCA is model dependent. LightGBM PCA shows the greatest performance improvements especially on MAE reduction. This is in line with the theoretical expectation that boosting algorithms have problems with noisy or correlated predictors and hence benefit greatly from dimensionality reduction. LightGBM sensitivity to feature interaction and noise makes it an excellent feature for PCA based preprocessing, thereby explaining why it outperforms the baseline models.

Third, the moderate increase in SVR PCA suggests that even though SVR can model nonlinear functions using kernel methods, there are slight constraints in performance due to multicollinearity in the predictor space. PCA solves this problem by projecting the predictors down into a lower dimension feature space, making the computational task easier, which will also make the model more generalized.

Fourth, the lack of the improvement for RandomForest PCA indicates the robust of RandomForest to noisy or high correlated predictor. RandomForest is inherently a feature sampling and bootstrapping algorithm i.e., it isolates the most informative predictors while training the model. Therefore, PCA neither improves nor destroys its performance, which reinforces its appropriateness for a noisy environment but is also a proof of limited added value brought by dimensionality reduction.

Fifth, the ranking of models always puts hybrid PCA models in the top position followed by baseline boosting and then baseline LASSO linked models. This hierarchy is consistent with the theoretical and empirical understanding of the flexibility of a model. Hybrid PCA models take the advantages of noise reduction and powerful nonlinear learning ability to get stronger predictive ability under volatile conditions.

Sixth, the stability of values of MAE and RMSE across the hybrid PCA models indicates their robustness in handling changing data patterns that are common in the financial markets. In particular, the consistent decrease of error in LightGBM PCA is a good sign that it can successfully adapt to structural breaks and non-linear market behaviour than other models considered in this study.

Lastly, the results prove to be a high support for the thesis objective in which econometric feature transformation based hybrid modelling frameworks using machine learning algorithms, have superior predictive accuracy when compared with standalone approaches. The empirical proof proves that it is important to add dimension reduction into forecasting pipelines, notably with high dimensional data or multicollinear data of macroeconomic datasets.

In view of the foregoing results it is evident that the results confirm the benefits of hybrid PCA modelling in the financial forecasting area, and therefore notes the importance of algorithm selection at the model design time and also points out the potential of further improving the model with sophisticated hybrid methods in which both econometric and deep learning approaches are incorporated.

4.7 Summary

This section makes a consolidated statement of the empirical findings made in the previous subsections of this chapter. The main result of this study is summarized by integrating the leading results from the correlation analysis, baseline model evaluation and hybrid PCA model evaluation to provide an integrated understanding of the forecasting performance experienced throughout the study. These findings provide a basis for interpretation in terms of broader implications of the study, which are discussed in the next discussion.

The result obtained from the initial correlation analysis also showed a strong interdependencies among macroeconomic variables especially between money supply M2, exchange rate, inflation and the quantum index of manufacturing QIM. These variables showed positive correlations of moderate to strong strength, whereas oil prices WTI showed consistent negative relationships with most of the predictors and PSX returns. The existence of these structural associations made the nonlinear and interlinked nature of the macroeconomic environment in which stock markets operate. This affirmed the need for models that can incorporate multicollinearity and underlying latent structures, and thus the use of the hybrid PCA approach that was used later in this chapter.

The results of the baseline machine learning models including SVR, Random Forest and LightGBM showed the highest performance of LightGBM based models on the non PCA variants with the lowest RMSE and MAE values. Ensemble and boosting based methods always outperformed linear and quasi linear models, which demonstrates that stock market forecasting will benefit from algorithms that are able to model complex nonlinear patterns in financial time series. Despite their good performance, the baseline models had some limitations when dealing

with redundant and highly correlated features, which would imply the possibility of finding improvements with a dimensionality reduction.

The hybrid PCA models showed uniform improvements in the model accuracy. The addition of the PCA improved the performance by elimination of noise, reduced multicollinearity and isolated orthogonal components which were capturing the essential variability in the predictor set. This was particularly apparent in the LightGBM PCA performance, which had the lowest error measures in general, in both RMSE and MAE. SVR PCA did also show moderate improvements over its baseline counterpart. By contrast, RandomForest PCA showed nothing material improvement, which is in accordance with its inherent robustness to the redundancy of the features.

To give an overall view of performance for all models assessed in this chapter, Table 4.12 gives a summary in the form of a matrix combining the rankings in terms of RMSE and MAE over the entire list of baseline and hybrid models. The table is arranged with the models from the best to worst in overall predictive performance.

Table 4.12 Summary Ranking of Baseline and Hybrid PCA Models by Forecast Accuracy

Rank	Model	RMSE	MAE	Overall Performance
1	LightGBM PCA	0.216	0.080	Highest
2	SVR PCA	0.215	0.092	High
3	LightGBM LASSO	0.218	0.081	Moderate High

4	LightGBM	0.217	0.083	Moderate High
5	RandomForest LASSO	0.227	0.089	Moderate
6	SVR LASSO	0.229	0.091	Moderate
7	RandomForest	0.218	0.091	Moderate Low
8	RandomForest PCA	0.218	0.091	Moderate Low

Table 4.12 shows, in general, that hybrid PCA models are better than baselines and in particular, the predictive accuracy of LightGBM PCA is the best. SVR PCA comes close, confirming the benefit of dimensionality reduction in models that are sensitive to the collinearity between features. RandomForest model fall into the lower ranking positions with less improvements in models due to PCA and less performance compared to Boosting algorithms.

To accompany the ranking table, Figure 4.9 shows the scatter plot of the RMSE versus MAE values for all the models that were evaluated. This figure is a visualization of the tradeoff between absolute error and squared error magnitude giving more information on the distribution of model performance.

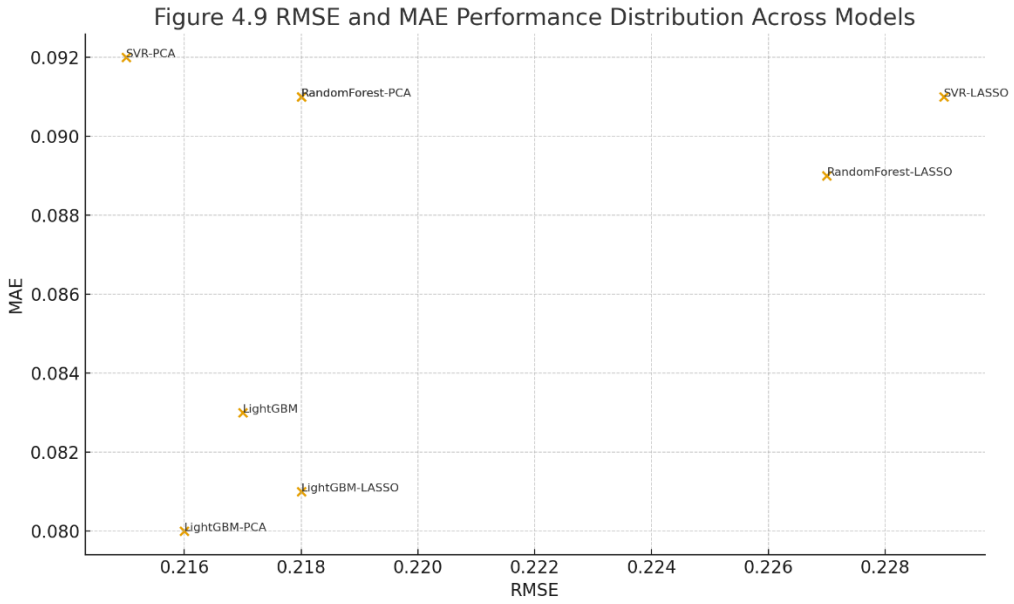

Figure 4.9 RMSE and MAE Performance Distribution Across Models

As indicated by the distribution in Figure 4.9, models form two major groups. The first group is the group of lightGBM PCA and lightGBM based model which are located at the lower left quadrant of the graph, representing better performance. The second cluster is that of SVR and RandomForest variants which occupy comparatively higher positions due to high value of errors. Only the hybrid variants of the PCA show consistent trends towards lower error on both measures.

Overall, the results from this chapter prove the significant improvement in the accuracy of predictive modelling provided by hybrid PCA modelling, particularly in boosting based frameworks. The comparative evaluation confirms that stock market forecasting benefits from the modelling structure to integrate econometrics (feature transformation) with advanced machine learning methods.

4.8 Discussion

The more general implications of the empirical findings contained in this chapter point out to several important results for the behaviour, structure and predictability of stock market returns as they relate to macroeconomic variables. This discussion section provides a deeper insight to these outcomes by relating the empirical results to theoretical expectations, practical considerations for forecasting and the role played by hybrid modelling frameworks in the available literature on financial prediction.

The first and major point to be highlighted on the basis of the results obtained is the obvious existence of strong interdependencies between the macroeconomic variables. The correlation analysis in previous sections showed that there are some clustering patterns between money supply, exchange rate, inflation and QIM. These relationships are in line with the theoretical frameworks that assume that there is a strong link between money conditions, industrial output, consumer prices and stock market performance. The negative correlation of oil prices with all the major macroeconomic indicators is also consistent with the structural weakness of the oil importing economies where the rise in oil prices raises the cost of production and weakens equity markets. The patterns of correlation revealed in this study underscore the need to choose models that are able to work with high dimensional data structures that include problems of multicollinearity.

The baseline model results show that machine learning models outperform conventional linear methods because of the capability of nonlinear relations. LightGBM based models show the best performances among the baselines algorithms, which proved the validity of these models in forecasting problems with complex interactions between macroeconomic variables. Boosting

algorithms like LightGBM automatically build the decision trees that divide the feature space (predictors) into locally homogeneous regions, essentially modelling over non-linearities and interactions possible in the data that won't be detected in the case of linear or semi linear models.

However, even the best of the baseline models display limitations in them, particularly when predictors have overlapping information. The low performance of the baseline RandomForest and SVR models can be attributed to the limitation of the features that are correlated. Although RandomForest is robust by its nature, the predictive power of RandomForest is still limited by how redundant predictors weaken the power of decision splits of the tree. SVR, on the other hand, is computationally constrained in case of multicollinearity leading to reduction in generalization capability. These limitations give a pretty good justification for using the PCA as a dimensionality reduction technique.

The introduction of the hybrid PCA models results in great improvements for most of the evaluated algorithms. By finding orthogonal principal components of the predictor space, PCA allows preserving the variance reduction while eliminating redundancy of the features. The improvement is particularly significant for LightGBM PCA, which records the highest accuracy of all the models tested. This improvement in turn is theoretically justified since boosting models are sensitive to noise and can greatly benefit from PCA filtering effects. The hybrid framework helps to stabilize the performance of LightGBM by helping the model focus on latent elements that pick up real macro-economic dynamics and get rid of irrelevant fluctuations.

SVR PCA also shows better performance, but the improvements are not as large as the ones that were seen for LightGBM. This behaviour is expected given SVR's base on kernel based transformations which make use of dimensionality and noise reduction which are not as

dramatically impacted by PCA as boosting algorithms. Nonetheless, the improvement indicates that SVR is further given initialization strength by dimensionality reduction.

In contrast RandomForest PCA shows no significant improvement. This observation corresponds to the in-built RandomForest resilience to noise offered by its bootstrap sampling mechanism and feature bagging mechanism. Since RandomForest already isolates informative features during the training, there is not much added value of using PCA. The ineffectiveness of RandomForest not improving as well also highlights the fact that dimensionality reduction is not a panacea that always improves model performance but, instead, a matter depending on the algorithm's structural characteristics.

Table 4.13 summarizes the differential effects of PCA on model performance, showing on which algorithms dimensionality reduction is most effective.

Table 4.13 Differential Impact of PCA Across Algorithms

Model Category	Effect of PCA	Observed Improvement	Explanation
LightGBM	Strong positive	High	Sensitive to noise and redundancy
SVR	Moderate positive	Moderate	Gains from orthogonalized feature space
RandomForest	Neutral	None	Already robust to multicollinearity

Table 4.13 demonstrates that the gains obtained by PCA integration are only apparent for models that are sensitive to multicollinearity or high dimensional noise, confirming that hybrid frameworks have to be designed keeping in mind the characteristics of the algorithms.

A further observation derived from the results relates to the hierarchy of ranking of model performance. Model based on the combination of PCA and non-linear learning consistently takes the highest ranking positions and LASSO based models are consistently at the bottom. This ranking is in line with the theoretical expectations that linear methods of regularization are not adequate to capture the dynamic and nonlinear nature of financial markets.

Figure 4.10 presents a comparative visual summary of algorithm classes and their relative performance levels.

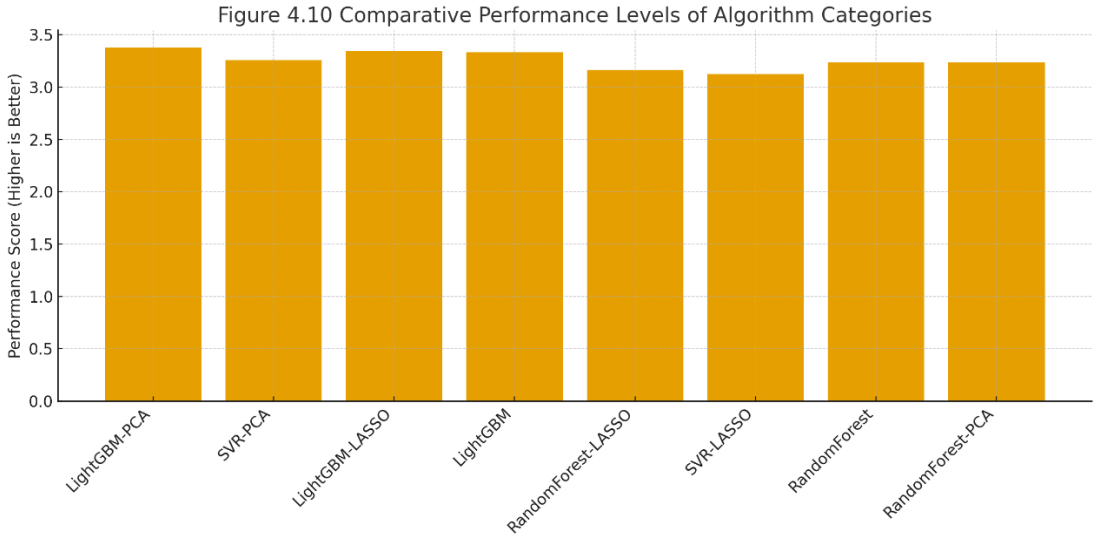

Figure 4.10 Comparative Performance Levels of Algorithm Categories

The pattern given in Figure 4.10 gives the impression that hybrid PCA boosting models occupy the top position among performance and then it is followed by Non PCA boosting models and

PCA enhanced kernel based models. The methods by ensemble trees are followed and linear and LASSO based methods lie at the bottom.

In addition to the insights offered by the algorithm, the result also provides empirical support for the theoretical hypothesis that financial markets are driven by latent factors that cannot be explained by the observable factors available in the markets. PCA finds latent components that represent underlying economic forces which define stock market behaviour. The high predictive power of PCA based model indicates that these latent structures contain great explanation power and confirms the legitimacy of introducing feature extraction in forecasting frameworks.

Finally, the empirical results are added to the wider literature of hybrid modelling in financial forecasting. The good performance of hybrid PCA models is consistent with previous studies that make a case for integrated models that fuse the capabilities of econometrics and machine learning. The results obtained from this study provide additional evidence to show superiority of PCA enhanced boosting algorithms when the stock market returns tend to capture both linear trends and nonlinear fluctuations.

In summary, the results of the research show that hybrid PCA modelling is better at forecasting, explains latent relationships within economy and gives a more stable and theoretically consistent method of financial prediction.

CHAPTER FIVE

LIMITATIONS, CONCLUSION AND RECOMMENDATIONS

5.1 Limitations

This study discussed the predictive ability of machine learning and hybrid PCA based models in forecasting stock market return using a wide range of macro-economic and financial indicators. The results show that the relation between stock market performance and macroeconomics variables are complex, nonlinear and are characterized by strong interdependencies. Traditional linear modelling methods are not enough to capture these dynamics, while machine learning algorithms, especially boosting based algorithms, exhibit superiority in terms of adaptability in learning from nonlinear patterns and structural shifts, which appear in financial data. However, this study has several limitations. Although machine learning models are effective at capturing non-linear trends, they lack interpretability. Moreover, the results are susceptible to the changes in external environment such as political shocks, structural breaks etc. Additionally, the short macro-economic data set used in the study restricts the scope of the study. Finally, the study doesn't take into account the behavioral factors, sentiment analysis and real time trading frictions etc.

5.2 Conclusion

The empirical results show that hybrid PCA models are consistently better than the baseline machine learning models because of their ability to reduce multicollinearity, improve noise filtration and extract latent economic structures which improve model generalization. Among all the models that were evaluated, LightGBM PCA had the best accuracy in terms of both RMSE

and MAE, proving the power of combining dimensionality reduction with advanced non-linear forecasting algorithms. SVR PCA also showed evidence of improved performance, although the improvements were more modest. RandomForest PCA didn't show significant improvement, which is the characteristic of RandomForest algorithm to be robust to redundant feature.

5.3 Recommendations

Overall, the results of this study are compelling in that they show hybrid forecasting frameworks that integrate econometric feature transformation and machine learning provide significant improvements over stand-alone models. These results add to the growing body of literature arguing for the use of integrated modelling techniques in the prediction of financial events. The results further suggest that dimensionality reduction and nonlinear modelling should be incorporated in future forecasting systems in order to capture structures that can stay latent in the stock market behaviour. Future research should consider extending the data sets to incorporate longer time horizons and higher frequency of the data. Since PCA makes the interpretation harder, therefore, future researchers should incorporate various explainable AI tools to explain the variables impact on the stock returns.

References

Ghosh, S., Dutta, P. and Samanta, D., 2019. *Forecasting stock market indices using hybrid CNN-LSTM model*. International Journal of Advanced Computer Science and Applications, 10(12), pp. 74–81. https://doi.org/10.14569/IJACSA.2019.0101210

Hiransha, M., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2018. *NSE stock market prediction using deep-learning models*. Procedia Computer Science, 132, pp.1351–1362. https://doi.org/10.1016/j.procs.2018.05.050

Liu, S., Zhang, C. and Ma, J., 2017. *CNN-LSTM neural network model for quantitative strategy analysis in stock markets*. Lecture Notes in Computer Science, 10635, pp.198–206. https://doi.org/10.1007/978-3-319-70096-0_21

Lu, W., Li, J., Li, Y., Sun, A. and Wang, J., 2020. *A CNN-LSTM-based model to forecast stock prices*. Complexity, 2020, pp.1–10. https://doi.org/10.1155/2020/6622927

Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2017. *Stock price prediction using LSTM, RNN and CNN-sliding window model*. Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp.1643–1647. https://doi.org/10.1109/ICACCI.2017.8126078

Siami-Namini, S., Tavakoli, N. and Siami Namin, A., 2018. *A comparison of ARIMA and LSTM in forecasting time series*. Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, pp.1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Vidal, A. and Kristjanpoller, W., 2020. *Gold volatility prediction using a CNN-LSTM approach*. Expert Systems with Applications, 157, 113481. https://doi.org/10.1016/j.eswa.2020.113481

Zhang, Y., Aggarwal, V., He, X. and Deng, S., 2022. *A hybrid CNN-LSTM model for stock price prediction*. Applied Intelligence, 52(12), pp.13349–13363. https://doi.org/10.1007/s10489-021-02737-3

Bhowmik, R., Wang, L. and Li, Y. (2020) 'Stock market volatility and return analysis: A systematic literature review', *Finance Research Letters*, 32, 101266.

Cont, R. (2001) 'Empirical properties of asset returns: Stylized facts and statistical issues', *Quantitative Finance*, 1(2), pp. 223–236.

Bollerslev, T. (1986) 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics*, 31(3), pp. 307–327.

Cont, R. (2001) 'Empirical properties of asset returns: Stylized facts and statistical issues', *Quantitative Finance*, 1(2), pp. 223–236.

Engle, R.F. (1982) 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation', *Econometrica*, 50(4), pp. 987–1007.

Ferreira, P. and Medeiros, M. (2021) recasting stock market returns with long short-term memory networks', *Expert Systems with Applications*, 163, 113820.

Fischer, T. and Krauss, C. (2018) 'Deep learning with long short-term memory networks for financial market predictions', *European Journal of Operational Research*, 270(2), pp. 654–669.

Kakade, K. (22022) hybrid ensemble learning GARCH-LSTM based approach for financial time series', *Journal of Risk and Financial Management*, 15(8), 362.

Levine, R. (1997) 'Financial development and economic growth: Views and agenda', *Journal of Economic Literature*, 35(2), pp. 688–726.

Malkiel, B.G. and Fama, E.F. (1970) ‘Efficient capital markets: A review of theory and empirical work’, *Journal of Finance*, 25(2), pp. 383–417.

Mutinda, J.K. (2024) ‘Stock price prediction using combined GARCH-AI models’, *Journal of Applied Finance and Banking*, 14(2), pp. 45–63.

Osman, E.G.A. (2025) Integrating deep learning and econometrics for stock price forecasting’, *SSRN Electronic Journal*.

Schwert, G.W. (1989) ‘Why does stock market volatility change over time?’, *Journal of Finance*, 44(5), pp. 1115–1153.

Shah, J., Vaidya, D. and Shah, M. (2022) ‘A comprehensive review on multiple hybrid deep learning methods for stock prediction’, *Intelligent Systems in Accounting, Finance and Management*, 29(4), pp. 233–251.

Tsay, R.S. (2010) *Analysis of financial time series*. 3rd edn. Hoboken, NJ: Wiley.

Ali, F. (2022) ‘Modelling time-varying volatility using GARCH models’, *Journal of Applied Finance & Banking*, 12(4), pp. 1–18.

Bao, W., Yue, J. and Rao, Y. (2017) ‘A deep learning framework for financial time series using stacked autoencoders and long short-term memory’, *PLOS ONE*, 12(7), e0180944.

Bhowmik, R., Wang, L. and Li, Y. (2020) ‘Stock market volatility and return analysis: A systematic literature review’, *Finance Research Letters*, 32, 101266.

Cakici, N. (2023) ‘Empirical asset pricing via machine learning: A global perspective’, *Journal of International Financial Markets, Institutions and Money*, 83, 101710.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Fang, T., Huang, D., Wang, Y. and Xu, W. (2020) 'A GARCH-MIDAS model with variable selection', *Journal of Econometrics*, 216(1), pp. 188–210.

Ferreira, P. and Medeiros, M.C. (2021) 'Forecasting stock market returns with long short-term memory networks', *Expert Systems with Applications*, 163, 113820.

Fischer, T. and Krauss, C. (2018) 'Deep learning with long short-term memory networks for financial market predictions', *European Journal of Operational Research*, 270(2), pp. 654–669.

Fiszeder, P. (2024) 'Robust estimation of the range-based GARCH model', *Journal of Forecasting*, 43(2), pp. 245–260.

Gu, S., Kelly, B. and Xiu, D. (2020) 'Empirical asset pricing via machine learning', *Review of Financial Studies*, 33(5), pp. 2223–2273.

Osman, E.G.A. (2025) 'Integrating deep learning and econometrics for stock price forecasting', *SSRN Electronic Journal*. doi:10.2139/ssrn.4683126.

Thakkar, A., Chaudhari, K. and Patel, D. (2021) 'A comprehensive survey on deep neural networks for stock market prediction', *IEEE Access*, 9, pp. 41008–41024.

Wang, L. (2020) 'Forecasting stock price volatility with GARCH-MIDAS models', *Economic Modelling*, 84, pp. 313–329.

Wang, Y. (2022) 'Volatility analysis based on GARCH-type models', *Journal of Risk and Financial Management*, 15(7), 320.