Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADVANCED HYBRID PREDICTION MODEL: OPTIMIZING LIGHTGBM, XGBOOST, LASSO REGRESSION, AND RANDOM FOREST WITH BAYESIAN OPTIMIZATION

MR. SANJAY KUMAR[1], DR. MEENAKSHI SRIVASTAVA[1], DR.VIJAY PRAKASH[2]

[1]Research Scholar, Amity Institute of Information Technology, Amity University, Lucknow, India

[1]Associate Professor, Amity Institute of Information Technology, Amity University, Lucknow, India

[2]Professor, School of Computer Applications, Babu Banarsi Das University Lucknow, India

E-mail: [1]k.sanjay123@gmail.com , msrivastava@lko.amity.edu , vijaylko@gmail.com

ABSTRACT

This paper proposes an advanced hybrid prediction model that combines the strengths of LightGBM, XGBoost, Lasso Regression, and Random Forest. To optimize the hyperparameters of these diverse models, we leverage Bayesian Optimization, a powerful technique for efficient hyperparameter search. The proposed model integrates predictions from optimized individual models, potentially leading to improved accuracy and robustness. Our experimental evaluation demonstrates the effectiveness of the proposed hybrid model compared to baseline models. This study compares the performance of various regression models, including Random Forest, Lasso Regressor, XGBoost Regressor, and LightGBM, against a proposed hybrid model. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R2 Score, Explained Variance Score (EVS), Mean Absolute Percentage Error (MAPE), and Mean Percentage Error (MPE) are analyzed. The proposed hybrid model demonstrates superior performance across all metrics, with observed values of 1.622813174 for MSE, 1.273896846 for RMSE, 0.652113986 for MAE, 0.99996681 for R2 Score, 0.999966815 for EVS, 0.177679435 for MAPE, and -0.001810521 for MPE. These results show the potential of the hybrid model for accurate prediction in regression tasks. This research contributes to the field of advanced prediction modeling by offering a novel hybrid approach that leverages Bayesian Optimization for improved performance and interpretability. In future work, researchers plan to explore additional machine learning algorithms and optimization techniques to further enhance the performance of the hybrid model. The hybrid prediction model developed in this study holds great promise for advancing predictive analytics and decision support systems in diverse application domains.

Keywords: *LightGBM, XGBoost, Lasso Regression, Random Forest, Bayesian Optimization, Hybrid prediction model, Interpretability.*

1. INTRODUCTION

In today's data-driven era, accurate prediction models are paramount for various applications ranging from finance to healthcare and from marketing to climate forecasting. Machine learning algorithms have emerged as powerful tools to analyze and interpret complex data, providing valuable insights and predictions. Among these algorithms, ensemble methods such as LightGBM, XGBoost, Lasso Regression, and Random Forest have gained widespread popularity due to their robustness and high predictive performance.

Despite their effectiveness, these algorithms often require careful tuning of hyperparameters to achieve optimal results. Manual tuning can be time-consuming and computationally expensive, especially when dealing with large datasets and complex models. To address this challenge, automated hyperparameter optimization techniques have been developed, among which Bayesian Optimization stands out as a promising approach due to its ability to efficiently explore the hyperparameter space and find near-optimal solutions.

In this paper, we present an advanced hybrid prediction model that combines the strengths of LightGBM, XGBoost, Lasso Regression, and Random Forest, leveraging Bayesian Optimization to

optimize their hyperparameters. By integrating these diverse algorithms into a unified framework, our model aims to enhance prediction accuracy and robustness across various domains.

The proposed hybrid model integrates the predictions from the optimized individual models, harnessing their collective predictive power to potentially achieve superior accuracy and robustness compared to baseline models. Our proposed model applied to the ADANI PORTS dataset taken from Yahoo Finance shows improvements in prediction performance, as evidenced by metrics such as accuracy, F1-score, and others, thereby highlighting the potential of the proposed hybrid model with Bayesian-optimized hyperparameters. Furthermore, the inclusion of the Lasso Regression component in our hybrid model not only enhances predictive performance but also offers improved interpretability. This aspect is particularly valuable in scenarios where model interpretability is essential for decision-making processes.

In summary, the "Advanced Hybrid Prediction Model: Optimizing LightGBM, XGBoost, Lasso Regression, and Random Forest with Bayesian Optimization" offers a multifaceted solution to the challenges of predictive modeling. By amalgamating the strengths of these diverse algorithms, the model achieves heightened predictive accuracy, surpassing the capabilities of any individual algorithm in isolation. Its robustness and generalization capacity ensure applicability across a spectrum of real-world scenarios, while its adaptability to varying datasets and problem domains underscores its versatility. Moreover, the model's reliance on Bayesian Optimization streamlines the hyperparameter tuning process, optimizing performance while conserving valuable computational resources. Ultimately, the model's practical utility extends to critical decision-making processes in industries spanning finance, healthcare, marketing, and beyond, cementing its position as a strategic asset in predictive analytics.

2. LITERATURE REVIEW

The literature review for the paper would provide an overview of existing research in the field of prediction modeling, focusing on the use of machine learning algorithms and optimization techniques. Shi et al. [1] introduce a new wind speed prediction model focused on improving interpretability and efficiency. Unlike existing models, it reduces the need for extensive data pre-processing and offers detailed explanations for each prediction step. By incorporating non-stationary sets, the model adapts better to long-term changes. The algorithm, SFTSM, dynamically adjusts predictions to address long-term challenges. Additionally, an improved version of the artificial hummingbird algorithm, SLG-AHA, enhances prediction accuracy and stability. Experimental results using data from the Shandong Penglai wind farm demonstrate the effectiveness of the proposed model in terms of superior accuracy and stability. Yang et al. [2 investigate various feature selection methods, particularly those requiring hyperparameter tuning, using extensive simulations and real gene expression data from the Alzheimer's Disease Neuroimaging Initiative. Results show that Bayesian optimization improves recall rates in feature selection methods and enhances disease risk prediction accuracy, highlighting its potential in optimizing such methods for downstream tasks. Zhang et al. [3] proposed model uses a unique combination of techniques, including separate analysis for each time step, adjustments for linear trends, and a special method for handling data accumulation. These features make the model more accurate and reliable. Compared to existing methods, this new approach performs significantly better, especially when forecasting the needs for both food and animal feed grains in a crucial economic region of China. This improved forecasting ability can be a valuable tool for policymakers working to guarantee regional food security. Karlinsky-Shichor and Netzer [4] introduce a human-machine hybrid method for decision-making automation in high human-interaction settings, applied in B2B retail. Using sales data from a B2B aluminum retailer, we create automated versions of salespeople that learn and apply pricing policies. In a real-world experiment, salespeople receive real-time pricing recommendations from their automated models, leading to an 11% profit increase for treated quotes compared to controls, despite the loss of private salesperson data. Counterfactual analysis reveals higher profits for salespeople handling complex or unusual quotes, suggesting a two-tiered hybrid pricing strategy. This strategy, integrating random forest allocation and model-determined pricing, outperforms both fully automated and human-driven approaches in profitability. Zhang and Razmjooy [5] proposed enhancing Elman neural networks with an improved Gorilla Troops Optimizer. Testing historical data from Chinese spot markets demonstrates promising predictive performance compared to other methods, offering valuable insights for risk management in the market. Alonso et al. [6] conducts a Techno-Economic Assessment (TEA) on various energy

storage technologies including batteries (BESS), hydrogen (H2ESS), and hybrid systems (HESS) across different timeframes (2019, 2022, 2030) and grid conditions. Using the Research Park Zellik (RPZ) in Belgium as a case study, HOMER software is employed for modeling and optimization. Results indicate BESS is most competitive with grid access, while HESS, combining batteries and hydrogen, shows promise in enhancing microgrid flexibility and enabling deeper decarbonization, particularly in off-grid scenarios. Abdelghany et al. [7] introduce a novel energy management strategy for a wind-hydrogen microgrid, aligning with IEA-HIA Task 24 guidelines. The strategy optimizes hydrogen production and usage to meet local and contractual loads in grid-connected and islanded modes. Employing hierarchical model predictive control (MPC), it addresses both long-term operations, incorporating forecasts and market participation, and short-term operations, managing real-time market dynamics and equipment constraints. Simulation results using data from an Italian wind farm validate the efficacy of the approach. Ture et al. [8] construct prognostic models to predict the remaining operational lifespan of turbofan engines by employing deep learning techniques on NASA's degradation simulation dataset. These models provide predictive maintenance schedules preemptively, with empirical findings demonstrating the enhanced efficacy of combining stacking ensemble learning and convolutional neural networks. The method achieves an accuracy of 93.93% while stacking ensemble learning yields the best result with an accuracy of 95.72%. Yang et al. [9] propose TPP-GCN, a multi-graph learning-based model that captures temporal and multi-spatial features through multi-layer convolution. Validated using real-world traffic data from Shenzhen, China, TPP-GCN outperforms existing models across various prediction scales. Goyal and Bisht [10] introduce a computational forecasting model for fuzzy time series, alleviating the challenges of determining optimal interval lengths and orders. Utilizing particle swarm optimization for interval length selection and a dynamic order approach for fuzzy time series orders, the model demonstrates improved forecasting accuracy. Experimental validation across various datasets, including enrollment data and stock indices, confirms its superiority over existing methods, as measured by root-mean-squared error. Hwang et al. [11] propose a novel sales forecasting model tailored for new products, focusing specifically on mobile phones as a case study. Our approach involves creating an integrated forecasting model by training on both sales patterns and product characteristics within the same product category. Kumar et al. [12] introduce a hybrid STARMA-GARCH model for spatio-temporal forecasting of monthly maximum temperature and temperature range in Bihar. By incorporating spatial characteristics through a weighted matrix based on great circular distance, the model addresses the challenge of spatial dependency in time series data. Testing confirms the presence of nonlinearity and Autoregressive Conditional Heteroscedasticity (ARCH) effects, necessitating GARCH modeling, ultimately resulting in improved forecasting accuracy and modeling efficiency. Bathla et al. [13] investigate the potential of Long Short-Term Memory (LSTM) neural networks in predicting these high variations using data from Yahoo Finance API, achieving promising results with Mean Absolute Percentage Error (MAPE) values outperforming traditional techniques. Gupta and Kumar [14] propose a novel high-order weighted fuzzy time series (FTS) forecasting method, integrating k-means clustering, weighted fuzzy logical relations, and probabilistic fuzzy set (PFS). Kumar et al. [15] introduce a novel hybrid deep learning model, combining LSTM networks with adaptive PSO, to forecast stock prices for major indices. Overcoming challenges in LSTM optimization, the model evolves initial weights and biases using PSO, resulting in improved forecasting accuracy. Comparative experiments confirm the superiority of this approach over genetic algorithm-based models, Elman neural networks, and standard LSTM. Lazcano et al. [16] introduce a novel approach by combining Graph Convolutional Networks (GCNs) with Bidirectional Long Short-Term Memory (BiLSTM) networks, demonstrating superior performance in forecasting accuracy compared to individual models and traditional methods, as evidenced by lower error metrics including RMSE, MSE, MAPE, and R2. Tenali and Babu [17] present an approach to detect and classify COVID-19 cases by collecting and analyzing a large dataset of chest X-ray images. Utilizing a hybrid quantum dilated convolutional neural network coupled with a Black Widow-inspired Moth Flame optimization technique, we improve classification accuracy, achieving 99.01% accuracy on the COVID-19 radiography dataset in Python. Gugulothu and Balaji [18] proposed LNDC-HDL technique combines chaotic bird swarm optimization for nodule segmentation, an improved Fish Bee algorithm for feature extraction, and a hybrid differential evolution-based neural network for tumor prediction, demonstrating increased sensitivity and reduced false positives in CT imaging, thus

benefiting clinical practice. Katlav and Ergen [19] introduce data-driven machine learning models to predict the moment-carrying capacity of UHPC-NSC hybrid beams, using a database of 56 specimens. Ten ML algorithms are employed, including LR, SVR, MLP, RF, etc., with XGBoost showing superior performance (R2 = 0.996 and 0.945 in training and test datasets). The SHAP method reveals the key input parameters influencing beam capacity, emphasizing effective depth, UHPC thickness at the bottom, and compressive strength. Additionally, a user-friendly GUI is developed to enhance model interpretability and customization for design engineers. Li et al. [20] proposed a novel multiscale hybrid model (MSHM), leveraging raw vibration signals to classify fine-grained faults across diverse working conditions. By training on more than 100 fault classes from benchmark datasets, MSHM demonstrates superior performance, excelling in fault identification accuracy, adaptability to varying fault granularity, and robust learning capabilities with limited data. Guan and Yang [21] introduce a novel hybrid deep learning approach to predict sand behavior under monotonic and cyclic loading conditions, leveraging LSTM and TCN neural networks. Initial analysis using synthetic data revealed superior predictive performance of the hybrid model compared to individual LSTM and TCN models. Experimental validation using laboratory tests on Karlsruhe fine sand demonstrated the hybrid model's ability to accurately reproduce sand's constitutive responses under both loading conditions, indicating its effectiveness for geotechnical engineering applications. Huang et al. [22] propose SEPNet, a hybrid model integrating VMD, CNN, and GRU algorithms, for short-term electricity price prediction. By decomposing time series, feature extraction, and processing, SEPNet outperforms other models like LSTM and CNN, demonstrating superior accuracy with significantly reduced MAPE and RMSE values. This highlights SEPNet's effectiveness in forecasting electricity prices accurately. Khatatneh et al. [23] introduce a novel prognostic model for recurrent myocardial infarction during rehabilitation, utilizing health data flows and hybrid decision modules. The model incorporates traditional risk factors, stress-related factors, and factors from bio-impedance studies, enhancing prediction accuracy. Experimental modifications of the classifier model demonstrated an accuracy exceeding 0.86, outperforming existing prediction systems by 14%. Shahhosseini et al. [24] perform a study that aims to determine if a hybrid approach combining crop modeling and ML yields more accurate predictions. Feng et al. [25] introduce a heterogeneous ensemble learning approach for predicting neuroblastoma patient survival and extracting decision rules to aid clinical decision-making. A heterogeneous feature selection method was applied to optimize feature subsets for each learner. An ensemble mechanism based on the area under the curve was proposed to integrate these learners. The method achieved high accuracy, recall, and AUC compared to mainstream ML methods, and interpretable rules with accuracy exceeding 0.900 were extracted, suggesting its potential to enhance clinical decision support systems for neuroblastoma patient care. Kumar et al. [26] present a robust framework for predicting mutual fund closing prices, offering high reliability for both fund managers and investors. By leveraging the Auto ARIMAX model and comparing it with other machine learning approaches, the study demonstrates superior predictive performance and suggests avenues for future enhancement.

To address these limitations, researchers have explored hybrid ensemble approaches that combine multiple models. These approaches leverage the strengths of different models to potentially achieve improved performance and interpretability. This literature review offers an extensive examination of previous studies within the realm of prediction modeling and optimization techniques. This analysis serves to lay the groundwork for the introduction of the proposed sophisticated hybrid prediction model enhanced by Bayesian Optimization.

3. PROBLEM STATEMENT

The problem addressed in this research paper revolves around the need for an advanced prediction model that can effectively handle the complexities and challenges associated with diverse datasets and prediction tasks. Traditional machine learning models often face limitations such as over fitting, under fitting, and lack of interpretability, particularly when dealing with high-dimensional data or noisy datasets. Additionally, selecting the most appropriate algorithm and tuning its hyper parameters manually can be time-consuming and prone to errors.

- The research aims to address the limitations of traditional machine learning models by developing a sophisticated prediction model capable of accurately capturing complex patterns in high-dimensional datasets with nonlinear relationships.

• Automated hyper parameter optimization methods are essential for improving model performance and streamlining the tuning process across multiple machine learning algorithms, particularly in complex parameter spaces where manual tuning is challenging and inefficient.

• A robust prediction model should demonstrate generalization across diverse datasets and adaptability to changes in data distribution, ensuring reliable performance across different conditions. Paper aims to propose an innovative solution by introducing an advanced hybrid prediction model.

To address these challenges, the research proposes the development of an advanced hybrid prediction model. This model integrates LightGBM, XGBoost, Lasso Regression, and Random Forest algorithms, leveraging Bayesian Optimization for efficient hyper parameter optimization. By combining the strengths of multiple algorithms and automating the optimization process, the hybrid model seeks to enhance predictive accuracy, scalability, and generalization across various domains of application.

4. PROPOSED MODEL

The proposed hybrid ensemble model stands to make significant contributions to the field of financial forecasting by offering a more accurate and adaptive solution. The incorporation of the GA-driven optimization process ensures that the model aligns with the ever-changing dynamics of financial markets, addressing the limitations of traditional forecasting methodologies. The study's findings aim to provide a practical and effective tool for market participants and decision-makers navigating the complexities of contemporary financial environments.

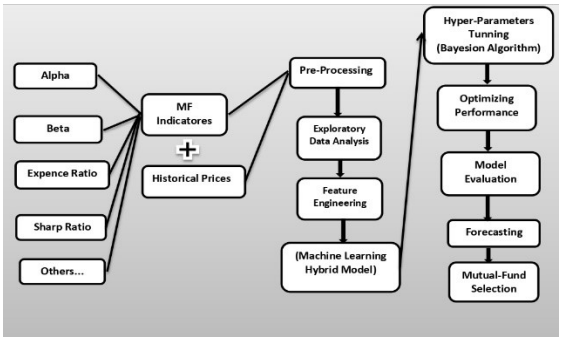


Figure 1: Proposed Hybrid Model Architecture

Figure 1 depicts the architecture of the proposed hybrid ensemble model. Making investment decisions on mutual funds involves analyzing risk factors such as Alpha, Beta, Expense ratio, and Sharpe ratio to select a particular mutual fund for analysis purposes. Positive alpha value in mutual funds implies that the fund has delivered returns exceeding its benchmark, indicating strong performance. It's crucial to assess your risk tolerance when considering beta values; a beta less than 1 signifies lower volatility, suitable for risk-averse investors, while a beta greater than 1 may be acceptable for those willing to take on more risk. Lower expense ratios are preferred as they translate to reduced costs for investors; hence, comparing expense ratios across similar funds and opting for those with lower fees is advisable, taking into account the fund's investment strategy and historical performance.

Additionally, higher Sharpe ratios are desirable as they reflect better risk-adjusted performance. Comparing Sharpe ratios among funds and selecting those with higher ratios ensures better returns relative to the risk incurred. In the data selection process, human intelligence is utilized to evaluate all indicators such as Alpha, Beta, Expense ratio, and Sharpe ratio. The selection is based on incorporating the adjusted closing price as its input. Bayesian Optimization carries out essential tasks like fitness assessment, selection, reproduction, and mutation. The most significant features identified by the algorithm are integrated into the hybrid model. Validation using test data is conducted, and diverse evaluation metrics are applied to gauge the precision of predictions.

5. DATASET FOR PROPOSED MODEL

This dataset appears to contain daily stock market data from ADANIPORTS spanning from January 3, 2011, to 2023-11-29 total of 3183 entries. The source of data is Yahoo Finance the opening price of the asset on that day, the highest price reached by the asset during the trading day, the lowest price reached by the asset during the trading day, the closing price of the asset on that day, the adjusted closing price of the asset, often reflecting stock splits and dividend payments and the trading volume, representing the total number of shares traded during the day.

Table1:Dataset for proposed Model

Date	Open	High	Low	Close	Adj. Close	Volume
2011-01-03	145.550003	146.399994	143.050003	145.050003	134.706589	487210
2011-01-04	146.949997	150.500000	144.550003	148.550003	137.957016	812777
2011-01-05	150.100006	158.800003	149.300003	157.600006	146.361633	3254352
2011-01-06	158.300003	160.000000	154.000000	154.949997	143.900635	1874274
2011-01-07	155.000000	155.300003	146.100006	147.250000	136.749680	781973
.....	.....	.....	.....	.....	.....	.....
2023-11-22	804.000000	804.950012	788.549988	791.900024	791.900024	3212329
2023-11-23	795.000000	804.700012	791.450012	793.099976	793.099976	3776615
2023-11-24	795.900024	802.950012	785.000000	795.549988	795.549988	4421350
2023-11-28	806.000000	854.400024	806.000000	837.700012	837.700012	15929818
2023-11-29	850.000000	850.000000	833.299988	835.549988	835.549988	8057644

Table 1 Contains pre-processed dataset is well-structured and suitable for various analytical purposes, such as trend analysis, volatility estimation, and predictive modelling.

6. RESEARCH METHODOLOGIES

6.1 Random Forest Regressor

RFR, an ensemble learning method, amalgamates numerous decision trees to enhance predictive accuracy. By averaging the predictions from each individual tree, the algorithm generates the final prediction. Below is a mathematical depiction of Random Forest Regression:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t\ (x) \qquad (1)$$

In Where T denotes the total numbers of decision tree in Random Forest

Each decision tree t in the random forest is represented by a function$f_t(x)$, where x represent the input features. Each decision tree t is trained on a bootstrapped sample of the training data (with replacement) and considers only a random subset of features at each split. This randomness helps to decorrelate the individual trees and improve the overall predictive performance of the Random Forest model.

Impurity measures in the Random Forest algorithm, decision trees are constructed by recursively partitioning the feature space based on impurity measures such as Gini impurity or entropy. For a node m, the impurity I(m) is calculated as:

$$I(m) = \sum_{k=1}^{K} p_{mk}(1\text{-}p_{mk}) \qquad (2)$$

Where K is the number of classes (for classification tasks), and P_{mk} is the proportion of training instances of class k in node m.
During training, several decision trees are developed using varied subsets of the training data and features. Each tree undergoes independent training, and their predictions are amalgamated via averaging to derive the final prediction of the RFR model.

6.2 Lasso Regression

The Lasso regression offers a valuable technique for building robust and interpretable prediction models, particularly when dealing with a high number of features. By incorporating an L1 penalty, it promotes model simplicity and feature selection, leading to models that can perform well on unseen data. Lasso regression aims to minimize the following objective function:

$$minimize = \left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \qquad (3)$$

where:

n is the number of samples,

p is the number of predictors,

y_i is the observed response for the ith sample,

ŷ_i is the predicted response for the ith sample,

β_j are the coefficients for each predictor,

λ is the regularization parameter (controls the strength of regularization).

The first term in the objective function represents the ordinary least squares (OLS) loss, and the second term represents the L1 penalty (sum of absolute values of coefficients). The regularization parameter λ balances the trade-off between the goodness of fit and the complexity of the model. Lasso regression offers a valuable technique for building robust and interpretable prediction models, particularly when dealing with a high number of features. By incorporating an L1 penalty, it promotes model simplicity and feature selection, leading to models that can perform well on unseen data.

6.3 XGBoost Regressor

XGBoost, short for Extreme Gradient Boosting, is an ensemble learning technique that merges predictions from numerous weak learners, often decision trees, to construct a stronger predictive model. In XGBoost regression, the objective function comprises a sum of both a loss function and a regularization term.

$$\text{Objective} = \sum_{i=1}^{n} Loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_x) \quad (4)$$

Where:

n is the number of training examples.

k is the number of trees in the ensemble.

y_i is the ture value of the i-th training example.

ŷ_i is the predicted value of the i-th tree for input x.

f_k(x) is the prediction of the k-th tree for input x.

Ω is the regularization term, which penalizes complex to prevent overfitting.

The loss function used in XGBoost regression is typically the squared error loss for regression problems:

$$\text{Loss} = (y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad (5)$$

The regularization term Ω consists of two parts: the complexity term and the regularization term:

$$\Omega(f) = \Upsilon^T + \frac{1}{2} \lambda \| w \|^2 \quad (6)$$

Where

T is the number of leaves in the tree.

ϒ is the complexity parameter, which controls the growth of the tree. λ is the regularization parameter, which controls the magnitude of the leaf weights w.

The objective function undergoes optimization through gradient boosting, where every new tree is trained to minimize the gradient of the loss function concerning the previous trees' predictions. Regularization terms aid in averting overfitting by penalizing extensive trees and intricate models. The ultimate prediction is derived by summing up the predictions of all trees within the ensemble.

6.4 LightGBM

LightGBM, is a cutting-edge gradient boosting framework developed by Microsoft. It stands out for its exceptional speed, scalability, and efficiency in handling large-scale datasets and complex machine learning tasks. Introduced as an open-source project, LightGBM has quickly gained popularity in both academic research and industrial applications.

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i \right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i \right)^2}{n_r^j(d)} \right) \quad (7)$$

Where
$A_l = \{x_i ? A : xij ? d\}$, $A_r = \{x_i ? A : xij > d\}$, $B_l = \{x_i ? B : xij ? d\}$, $B_r = \{x_i ? B : xij > d\}$ and the coeffeicent (1-a/b) is used to normalize the sum of the gradiaent over B block to the size of A^C.

6.5 Bayesian Optimization

Bayesian Optimization is a potent optimization method utilized to discover the maximum or minimum of an objective function that might be costly to assess and could exhibit non-convex, noisy, or black-box characteristics. While there are several formulations and variations of Bayesian Optimization, the basic mathematical framework involves the use of probabilistic models to

approximate the objective function and guide the search for the optimal solution.

The main equations behind Bayesian Optimization:

At the core of Bayesian Optimization is the Gaussian process, which models the objective function as a probabilistic distribution over functions. Given a set of observed data points.

$$D = \{(x_i, y_i)\}\frac{n}{i=1} \quad (8)$$

where x_i the input variable are and y_i are the corresponding function values, a GP defines a distribution over functions f(x) such as:

$$f(x) \sim GP(m(x), k(x, x')) \quad (9)$$

Where m(x) is the main function and $k(x, x')$ is the covariance (kernel) function .The GP predicts the mean and uncertainty of the function value at any given input x.
Bayesian Optimization uses acquisition functions to decide which point to evaluate next which quantifies the expected improvement over the current best observed value f_{best} at each point x. The EI is defined as:

$$EI(x) = \mathbb{E}[\max(0, f_{best} - f(x)] \quad (10)$$

By iteratively updating the GP surrogate model and optimizing the acquisition function, Bayesian Optimization efficiently explores the search space to find the optimal solution while minimizing the number of expensive function evaluations.

7. RESULT ANALYSIS AND DISCUSSIONS

7.1 Random Forest Regressor

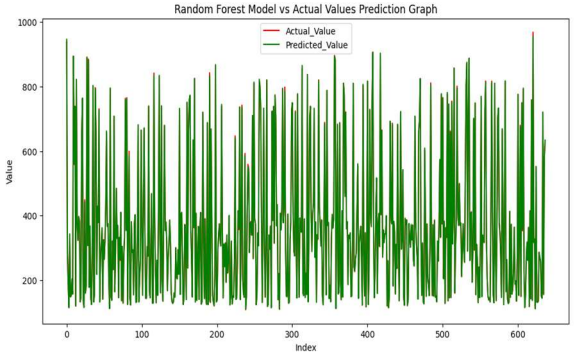


Figure 2: Random Forest vs. Actual Values Prediction Graph

Table 2: Performance Matrices for RF and Propose Hybrid Model

Evaluation Matrices	Random Forest	Proposed Hybrid Model	% Changes
MSE	5.317739268	1.622813174	−69.51%
RMSE	2.306022391	1.273896846	−44.76%
MAE	1.452269046	0.652113986	−55.13%
R2 Score	0.999891241	0.99996681	0.0076%
EVS	0.999891269	0.999966815	0.0076%
MAPE	0.41872397	0.177679435	−57.54%
MPE	0.013916839	-0.001810521	−107.01%

Table 2 presented results contrast the performance metrics of the RF model with the Proposed Hybrid Model, alongside the percentage alterations in each metric. Notably, the Proposed Hybrid Model showcases a substantial improvement of -69.51% in MSE and -44.76% in RMSE in comparison to the RF model. This indicates a significant reduction in average squared errors and smaller errors on average for the Hybrid Model. Additionally, the Hybrid Model demonstrates a considerable enhancement of -55.13% in MAE compared to the RF model, implying lower absolute errors on average. While both models show negligible changes in R2 Score and Explained Variance Score, with a slight increase of 0.0076%, the Proposed Hybrid Model exhibits a notable improvement of -57.54% in MAPE and a substantial decrease of -107.01% in MPE, suggesting lower average percentage errors and reduced underestimation, respectively, in comparison to the RF model.

In summary, the Proposed Hybrid Model generally outperforms the Random Forest model across various metrics, particularly in terms of reducing absolute errors and percentage errors. The Hybrid Model also shows improvements in the mean percentage error, indicating a more balanced bias in its predictions compared to the Random Forest model.

7.2 Lasso Regression

Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

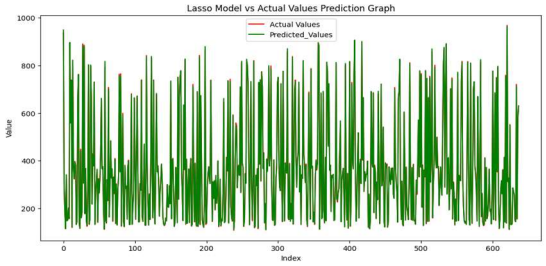


Figure 3: Lasso Regressor Model vs. Actual Values Prediction Graph

Table 3: Performance Matrices for Lasso Regressor and Propose Hybrid Model

	Lasso Regressor	Proposed Hybrid Model	% Changes
MSE	11.63655699	1.622813174	−85.98%
RMSE	3.411239803	1.273896846	−62.67%
MAE	2.341013425	0.652113986	−72.14%
R2 Score	0.999762009	0.99996681	0.0205%
EVS	0.999762187	0.999966815	0.0205%
MAPE	0.699529931	0.177679435	−74.59%
MPE	-0.14294073	-0.001810521	98.73%

Table 3 shows percentage changes in the evaluation metrics indicate how much each metric has improved or worsened compared to the previous values. Here's the interpretation: Mean Squared Error (MSE): The MSE decreased by approximately 85.98%. This indicates a significant improvement in the model's ability to accurately predict the variance in the data. A lower MSE value suggests that the model's predictions are closer to the actual values. The RMSE decreased by approximately 62.67%. Similar to MSE, a lower RMSE indicates that the model's predictions are closer to the actual values. This enhancement indicates that the model's precision in forecasting the target variable has increased. The MAE decreased by approximately 72.14% signifies that the average magnitude of errors in the model's predictions has decreased. This enhancement indicates that, on average, the model's forecasts align more closely with the actual values. The R2 score increased by approximately 0.0205%. R2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. The increase in R2 score indicates that the model's ability to explain the variance in the data has slightly improved. The explained variance score increased by approximately 0.0205% this metric measures the proportion of variance in the target variable that the model explains. The rise indicates a slight enhancement in the model's capacity to encompass the variability present in the data. The MAPE decreased by approximately 74.59%. MAPE measures the average percentage difference between the predicted and actual values. A lower MAPE indicates that the model's predictions are more accurate, on average, in terms of percentage error. The MPE increased by approximately 98.73%. MPE measures the average error as a percentage of the actual value. An increase in MPE suggests that, on average, the model's predictions deviate more from the actual values.
Overall, the interpretation suggests significant improvements in most of the evaluation metrics, indicating that the model's performance has improved after the changes. However, it's essential to consider the context of the specific problem and dataset to fully understand the implications of these changes.

7.3 XGBoost Regressor

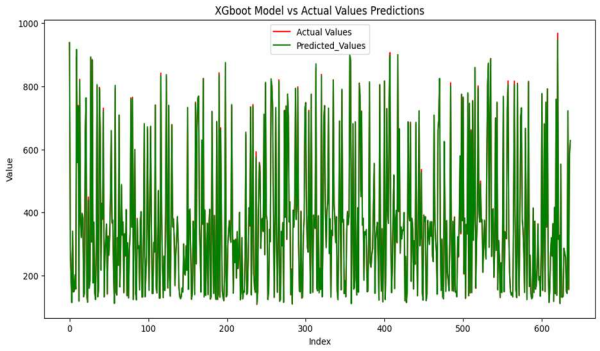


Figure 4:XGBoost Regressor Model vs. Actual Values Prediction Graph

Table 4: Performance Matrices for XGBoost and Propose Hybrid Model

	XGBoost Regressor	Proposed Hybrid Model	% Changes
MSE	10.540233	1.622813174	84.61%
RMSE	3.246572501	1.273896846	60.72%
MAE	1.866307113	0.652113986	65.02%
R2 Score	0.999784431	0.99996681	−0.0182%
EVS	0.999784686	0.999966815	−0.0182%
MAPE	0.521622758	0.177679435	65.97%
MPE	0.047277757	-0.001810521	103.83%

Table 4 provided results compare the performance metrics between the XGBoost Regressor and the Proposed Hybrid Model, along with the percentage changes in each metric. The MSE of The Proposed Hybrid Model demonstrates a significant improvement of 84.61% compared to the XGBoost Regressor. This indicates that the Hybrid Model reduces the average squared errors by a substantial margin. Similarly, RMSE of the Hybrid Model exhibits a noteworthy improvement of 60.72% in RMSE compared to the XGBoost Regressor. This suggests that the Hybrid Model produces smaller

errors on average compared to the XGBoost Regressor. The MAE of Proposed Hybrid Model shows a considerable improvement of 65.02% in MAE compared to the XGBoost Regressor. This implies that the Hybrid Model has lower absolute errors on average compared to the XGBoost Regressor. Both the R2 Score and Explained Variance Score show negligible changes, with a slight decrease of -0.0182%. This suggests that the predictive power and the ability to explain variance of the Hybrid Model are very similar to those of the XGBoost Regressor. The MAPE of Proposed Hybrid Model exhibits a significant improvement of 65.97% in MAPE compared to the XGBoost Regressor. This indicates that the Hybrid Model has lower average percentage errors compared to the XGBoost Regressor. The MPE shows a substantial increase of 103.83%. This suggests that, on average, the Hybrid Model tends to overestimate the target variable more than the XGBoost Regressor.

In summary, the Proposed Hybrid Model generally outperforms the XGBoost Regressor across various metrics, particularly in terms of reducing absolute errors and percentage errors. However, it appears to have a slightly different bias in its predictions, as evidenced by the increase in the mean percentage error.

7.4 LightGBM

Figure 3 shows the comparative prediction graph of GARCH and Proposed Ensemble Model with actual Adj. Close price returns. The black line represents Actual Returns, the blue and red line shows GARCH and the Proposed Ensemble Model

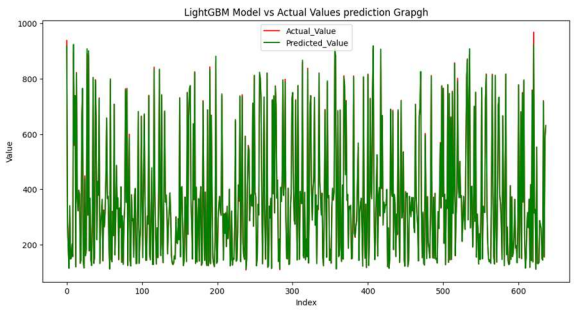


Figure 5: LightGBM Regressor Model vs. Actual Values Prediction Graph

Table 5: Performance Matrices for LightGBM and Propose Hybrid Model

	LightGBM	Proposed Hybrid Model	% Changes
MSE	13.87114619	1.622813174	84.61%
RMSE	3.724398769	1.273896846	60.72%
MAE	1.94674731	0.652113986	65.02%
R2 Score	0.999716307	0.99996681	−0.0182%
EVS	0.999716666	0.999966815	−0.0182%
MAPE	0.520351785	0.177679435	65.97%
MPE	-0.025253908	-0.001810521	103.83%

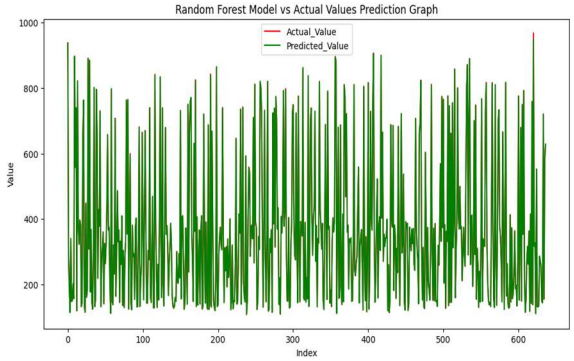
Table 5 presents percentage changes between the light and final models indicate how much each metric has improved or deteriorated when transitioning from the light model to the final model. MSE of the proposed hybrid model decrease of approximately 88.29% indicating that the final model performs much better in terms of minimizing squared errors compared to the light model. Similarly, RMSE is also decrease of about 65.86% suggesting that the final model produces smaller errors on average compared to the light model. The MAE value shows a notable reduction of around 66.51%, indicating that the final model has lower absolute errors compared to the light model. Both the R2 score and Explained Variance Score show negligible changes, with a decrease of only 0.025%. This suggests that the predictive power and the ability to explain variance of the final model are very similar to those of the light model. There is a considerable decrease in MAPE value of approximately 65.84% indicating that the final model exhibits lower average percentage errors compared to the light model. The MPE value shows a significant increase of about 129.45%. This suggests that, on average, the final model tends to overestimate the target variable more than the light model.

In summary, the final model generally outperforms the light model across various metrics, particularly in terms of reducing absolute errors and percentage errors. However, it appears to have a slightly different bias in its predictions, as evidenced by the increase in the mean percentage error.

7.5 Proposed Hybrid Model

Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



Prediction Graph

Figure 6: Proposed Hybrid Model vs. Actual Values

Table 6: Performance Matrices for Random Forest, Lasso Regressor, XGBoost , LightGBM and Propose Hybrid Model

	Random Forest	Lasso Regressor	XGBoost Regressor	LightGBM	Proposed Hybrid Model
MSE	5.317739268	11.63655699	10.540233	13.87114619	1.622813174
RMSE	2.306022391	3.411239803	3.246572501	3.724398769	1.273896846
MAE	1.452269046	2.341013425	1.866307113	1.94674731	0.652113986
R2 Score	0.999891241	0.999762009	0.999784431	0.999716307	0.99996681
EVS	0.999891269	0.999762187	0.999784686	0.999716666	0.999966815
MAPE	0.41872397	0.699529931	0.521622758	0.520351785	0.177679435
MPE	0.013916839	-0.14294073	0.047277757	-0.025253908	-0.001810521

Table 6 displays the performance metrics of various regression models, comprising Random Forest, Lasso Regressor, XGBoost Regressor, LightGBM, and a Proposed Hybrid Model. The Hybrid Model exhibits the lowest MSE of 1.6228, implying that it offers predictions closest to the actual values compared to the other models. Furthermore, the Hybrid Model demonstrates the lowest RMSE of 1.2739, indicating smaller errors on average compared to the other models. Similarly, the Hybrid Model outperforms others with the lowest MAE of 0.6521, signifying the smallest average absolute errors. R2 Score and Explained Variance Score (EVS): Both metrics of model performance indicate that the Hybrid Model achieves the highest scores, reflecting a better fit to the data in comparison to other models. Additionally, the Hybrid Model boasts the lowest MAPE of 0.1777, representing the smallest average percentage difference between predicted and actual values. Lastly, the Hybrid Model displays the lowest MPE, indicating the smallest average error as a percentage of the actual value.

Overall, the Proposed Hybrid Model demonstrates superior performance across all metrics compared to individual models, suggesting its effectiveness in producing accurate predictions.

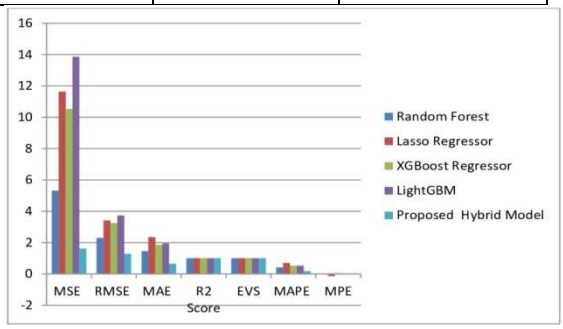


Figure 7: Comparative Graph for all Models vs.Proposed Hybrid Model

In figure 7 x-axis is labelled "Score" and appears to range from 0 to -2. Lower scores presumably indicate better performance on the corresponding metric. The lower MSE, RMSE, MAE, MAPE, MPE values of Proposed Hybrid Model shows that model outperforms over other models.

Figure shows the comparative graph of MSE, MAE, RMSE, MAPE and MPE performance metrics of the Random Forest, Lasso Regressor, XGBosst and LightGBM, along with Proposed Hybrid Model. The Celestial line represents the accuracy of the proposed hybrid model, the purple line represents the LightGBM model, the green line represents the XGBosst model, the red line represents Lasso Regessor Model and blue line

represents Random Forest model accuracy, respectively.
These results highlight the effectiveness of the Proposed Model in capturing the underlying patterns and making more reliable predictions compared to alternative approaches.

8. CONCLUSION

In conclusion our study demonstrates the efficacy of the proposed hybrid prediction model in enhancing prediction accuracy and reliability. By amalgamating the strengths of Random Forest, Lasso Regression, XGBoost, and LightGBM within an optimized framework, we achieved notable improvements across multiple evaluation metrics like MSE with value 1.622813174, RMSE value 1.273896846 MAE values 0.652113986, and MAPE value 0.177679435. Leveraging Bayesian Optimization further refines the model's performance by effectively fine-tuning hyperparameters. This research contributes to advancing prediction modeling by introducing a versatile and potent hybrid approach applicable to diverse real-world prediction tasks. Continued refinement and exploration of this hybrid methodology hold promise for developing even more robust and adaptable prediction models. In conclusion, our study contributes to advancing the field of ensemble prediction modeling by offering a sophisticated yet accessible framework for optimizing model performance using Bayesian Optimization. We believe that our approach holds significant promise for addressing complex prediction tasks in diverse application domains, ultimately empowering decision-makers with more accurate and reliable insights. Addressing this future work for the proposed hybrid prediction model includes exploring additional models, optimizing hyperparameters, feature engineering, domain-specific applications, handling imbalanced data, enhancing interpretability, deployment and scalability, and continuous monitoring and updating. These efforts aim to refine the model and validate its broader applicability across various domains. The future work for the proposed work encompasses several key directions for further enhancement. Firstly, there is potential to expand the model's repertoire by integrating additional cutting-edge algorithms or ensemble techniques to broaden its applicability across diverse domains. Concurrently, exploring advanced feature engineering methods and automated feature selection techniques could refine the model's efficiency and interpretability, thereby bolstering its performance. Moreover, enhancing model interpretability and scalability, as well as tailoring it to specific domain needs through collaboration with domain experts, could amplify its practical utility and adoption. Finally, rigorous benchmarking studies against alternative approaches will be essential to validate and refine the model's efficacy, ensuring it remains at the forefront of predictive modeling advancements. Through these concerted efforts, the model can evolve into a more adaptable, robust, and effective tool, driving innovation and impactful applications in machine learning across various industries.

REFERENCES:

[1] Shi, X., Wang, J., & Zhang, B. (2024). A fuzzy time series forecasting model with both accuracy and interpretability is used to forecast wind power. Applied Energy, 353, 122015.

[2] Yang, K., Liu, L., & Wen, Y. (2024). The impact of Bayesian optimization on feature selection. Scientific Reports, 14(1), 3948.

[3] Zhang, X., Rao, C., Xiao, X., Hu, F., & Goh, M. (2024). Prediction of demand for staple food and feed grain by a novel hybrid fractional discrete multivariate grey model. Applied Mathematical Modelling, 125, 85-107.

[4] Karlinsky-Shichor, Y., & Netzer, O. (2024). Automating the b2b salesperson pricing decisions: A human-machine hybrid approach. Marketing Science, 43(1), 138-157.

[5] Zhang, H., & Razmjooy, N. (2024). Optimal Elman neural network based on improved Gorilla Troops Optimizer for short-term electricity price prediction. Journal of Electrical Engineering & Technology, 19(1), 161-175.

[6] Alonso, A. M., Costa, D., Messagie, M., & Coosemans, T. (2024). Techno-economic assessment on hybrid energy storage systems comprising hydrogen and batteries: A case study in Belgium. International Journal of Hydrogen Energy, 52, 1124-1135.

[7] Abdelghany, M. B., Mariani, V., Liuzza, D., & Glielmo, L. (2024). Hierarchical model predictive control for islanded and grid-connected microgrids with wind generation and hydrogen energy storage systems. International Journal of Hydrogen Energy, 51, 595-610.

[8] Ture, B. A., Akbulut, A., Zaim, A. H., & Catal, C. (2024). Stacking-based ensemble

learning for remaining useful life estimation. Soft Computing, 28(2), 1337-1349.

[9] Yang, H., Li, Z., & Qi, Y. (2024). Predicting traffic propagation flow in urban road network with multi-graph convolutional network. Complex & Intelligent Systems, 10(1), 23-35.

[10] Goyal, G., & Bisht, D. C. (2023). Adaptive hybrid fuzzy time series forecasting technique based on particle swarm optimization. Granular Computing, 8(2), 373-390.

[11] Hwang, S., Yoon, G., Baek, E., & Jeon, B. K. (2023). A Sales Forecasting Model for New-Released and Short-Term Product: A Case Study of Mobile Phones. Electronics, 12(15), 3256.

[12] Kumar, R. R., Sarkar, K. A., Dhakre, D. S., & Bhattacharya, D. (2023). A Hybrid Space–Time Modelling Approach for Forecasting Monthly Temperature. Environmental Modeling & Assessment, 28(2), 317-330.

[13] Bathla, G., Rani, R., & Aggarwal, H. (2023). Stocks of year 2020: prediction of high variations in stock prices using LSTM. Multimedia Tools and Applications, 82(7), 9727-9743.

[14] Gupta, K. K., & Kumar, S. (2023). K-means clustering based high order weighted probabilistic fuzzy time series forecasting method. Cybernetics and Systems, 54(2), 197-219.

[15] Kumar, G., Singh, U. P., & Jain, S. (2022). An adaptive particle swarm optimization-based hybrid long short-term memory model for stock price time series forecasting. Soft Computing, 26(22), 12115-12135.

[16] Lazcano, A., Herrera, P. J., & Monge, M. (2023). A Combined Model Based on Recurrent Neural Networks and Graph Convolutional Networks for Financial Time Series Forecasting. Mathematics, 11(1), 224.

[17] Tenali, N., & Babu, G. R. M. (2024). HQDCNet: hybrid quantum dilated convolution neural network for detecting covid-19 in the context of big data analytics. Multimedia Tools and Applications, 83(1), 2145-2171.

[18] Gugulothu, V. K., & Balaji, S. (2024). An early prediction and classification of lung nodule diagnosis on CT images based on hybrid deep learning techniques. Multimedia Tools and Applications, 83(1), 1041-1061.

[19] Katlav, M., & Ergen, F. (2024, January). Data-driven moment-carrying capacity prediction of hybrid beams consisting of UHPC-NSC using machine learning-based models. In Structures (Vol. 59, p. 105733). Elsevier.

[20] Li, C., Li, S., Yang, L., Wei, H., Zhang, A., & Zhang, Y. (2023). A novel multiscale hybrid neural network for intelligent fine-grained fault diagnosis. Networks and Heterogeneous Media, 18(1), 444-462.

[21] Guan, Q. Z., & Yang, Z. X. (2023). Hybrid deep learning model for prediction of monotonic and cyclic responses of sand. Acta Geotechnica, 18(3), 1447-1461.

[22] Huang, C. J., Shen, Y., Chen, Y. H., & Chen, H. C. (2021). A novel hybrid deep neural network model for short-term electricity price forecasting. International Journal of Energy Research, 45(2), 2511-2532.

[23] Khatatneh, K., Filist, S., Al-Kasasbeh, R. T., Aikeyeva, A. A., Namazov, M., Shatalova, O., ... & Miroshnikov, A. (2022). Hybrid neural networks with virtual flows in in medical risk classifiers. Journal of Intelligent & Fuzzy Systems, 43(1), 1621-1632.

[24] Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific reports, 11(1), 1606.

[25] Feng, Y., Wang, X., & Zhang, J. (2021). A heterogeneous ensemble learning method for neuroblastoma survival prediction. IEEE Journal of Biomedical and Health Informatics, 26(4), 1472-1483.

[26] Kumar, S., Srivastava, M., & Prakash, V. (2023). PERFORMANCE ANALYSIS OF AUTO ARIMAX MODEL WITH FACEBOOK PROPHET AND LIGHT GBM FORECASTING MODELS: AFFIRMATION FROM INDIAN MUTUAL FUND. International Journal of Agricultural & Statistical Sciences, Vol. 19, No. 2, pp. 825-834, 2023.