OPTIMIZING PREDICTIVE ACCURACY WITH GRADIENT BOOSTED TREES IN FINANCIAL FORECASTING

Sathishkumar Chintala[a], Madan Mohan Tito Ayyalasomayajula[b]

[a]Independent Researcher, Texas, USA

[b]Computer Science, School of Business & Technology, Aspen University, Pheonix, AZ, USA

*Corresponding Author: Madan Mohan Tito Ayyalasomayajula

Abstract: Financial forecasting is indispensable for various financial sectors' strategic decision-making, risk management, and investment planning. Traditional statistical methods, though valuable, frequently struggle to capture the intricate, non-linear patterns embedded within financial data. In response to these limitations, Gradient Boosted Trees (GBTs) have emerged as a formidable ensemble learning technique renowned for enhancing predictive accuracy. This comprehensive review paper delves into applying GBTs in financial forecasting, scrutinizing their methodological underpinnings, distinct advantages, and comparative performance against other machine learning approaches. GBTs, with their real-world applications in financial forecasting, operate by sequentially combining multiple weak learners, typically decision trees, to iteratively refine predictions by focusing on residual errors. This iterative approach enables GBTs to effectively model complex relationships and interactions in financial datasets, outperforming traditional models like ARIMA and linear regression in many scenarios. Moreover, the paper addresses critical implementation challenges associated with GBTs, such as hyperparameter tuning and computational complexity, which are pivotal for achieving optimal performance. The review identifies promising avenues for future research, including integrating GBTs with deep learning techniques and advancements in real-time forecasting capabilities. By elucidating these aspects, this paper aims to provide insights that enhance the application and efficacy of GBTs in financial forecasting contexts.

Keywords: Gradient Boosting Trees, Financial Forecasting, Complex Relationships, Non-Linearity Handling, Overfitting Risk, Interpretability, Data Quality Sensitivity, Scalability, Computational Intensity, Parameter Sensitivity, XGBM, Lightgbm, Volatility Estimation.

1. Introduction

Financial forecasting is essential in economics, leading vital monetary policy decisions, influencing strategic investments, and strengthening effective risk management methods. The accuracy of prognoses is essential in calculating financial incentives and maintaining well-informed decision-making processes across sectors. Since their origin, forecasting approaches have relied on conventional statistical methods such as Autoregressive Integrated Moving Average (ARIMA) (Bakar & Rosbi, 2017) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) (Ilbeigi et al., 2017) models. Nonetheless, they have limits when dealing with the non-linear interrelationships common in financial data. The development of machine learning (ML) techniques has changed the landscape of financial forecasting, providing new tools for managing massive volumes of data and uncovering subtle patterns that traditional approaches lack. Gradient-boosted Trees (GBTs) are well-known among machine learning algorithms for their extraordinary ability to significantly improve forecast accuracy. As members of the ensemble learning family, GBTs build strong models by aggregating many weak learners, allowing them to correct previous trees' errors and improve overall performance.

2. Significance Of the Study

GBTs provide versatility in dealing with a wide range of target variables, whether continuous or categorical, and may support multidimensional feature relationships. Their versatility makes them a good fit for financial forecasting applications where understanding the underlying data structure and correctly predicting linkages is critical. Using GBTs, analysts may uncover hidden patterns and acquire insights into market behavior, resulting in more trustworthy and actionable projections. In addition, GBTs have the unique capacity to include external inputs, such as news sentiment and macroeconomic indices, into their prediction models. This adaptability allows them to capture the influence of exogenous factors on financial outcomes, increasing the model's overall explanatory power and boosting confidence in the provided projections. Thus, GBTs are an effective alternative for financial forecasters looking to manage the complexities of financial data and make accurate, intelligent forecasts. Also, GBTs' interpretability distinguishes them from other black-box ML models, such as deep neural networks. Their transparency allows for a better understanding of the underlying causes of financial patterns, permitting users to make educated choices based on explicit reasoning rather than depending exclusively on opaque outputs. This quality benefits financial forecasting, where openness and accountability are critical to effective decision-making. Continuous advances in processing resources and algorithmic breakthroughs broaden the usefulness and efficiency of GBTs in financial forecasting. These advancements allow more considerable information to be analyzed more successfully, opening new prospects for uncovering previously unnoticed patterns and delivering highly exact predictions. With continuing invention and refinement, GBTs have the potential to remain a powerful competitor in the search for accurate and insightful financial projections.

3. Evolution of Gradient Boosted Trees (GBTs)

GBTs have gained popularity in various fields, including finance, healthcare, and marketing, owing to their resilience and flexibility in processing complicated datasets with various attributes. One of the primary advantages of GBTs is their capacity to capture non-linear correlations between characteristics and target variables, which makes them preferable to simpler linear models in many real-world situations. GBTs are taught by incrementally adding decision trees, each trained to correct mistakes caused by combining all prior trees. This iterative procedure continues until a certain number of trees are achieved or no more improvement is possible. Gradient boosting, the approach that underpins GBTs, employs gradient descent optimization to minimize the loss function, guaranteeing that each new tree is added in a manner that reduces overall prediction error (Agapitos et al., 2017).

While strong, GBTs are susceptible to overfitting if not correctly calibrated. Regularization and cross-validation techniques are often used to avoid overfitting and guarantee that the model generalizes effectively to new data sets. Furthermore, advances in processing power and algorithmic optimizations have enabled practical training of GBTs on large-scale datasets, resulting in their widespread use in big data analysis. In recent years, various versions and upgrades to GBTs have arisen, including XGBoost, LightGBM, and CatBoost, each with advantages like quicker training speed, better handling of categorical variables, and higher accuracy. These versions have increased the applicability of GBTs to even more complicated problems, cementing their status as one of the most influential and adaptable machine-learning algorithms available today. From a technical aspect, the interpretability of GBTs varies according to the model's complexity and tree depth. Individual trees are easy to analyze, but the ensemble structure of GBTs complicates the interpretation of the whole model. Feature significance analysis and partial dependency plots are standard techniques for determining the relative value of various characteristics in prediction.

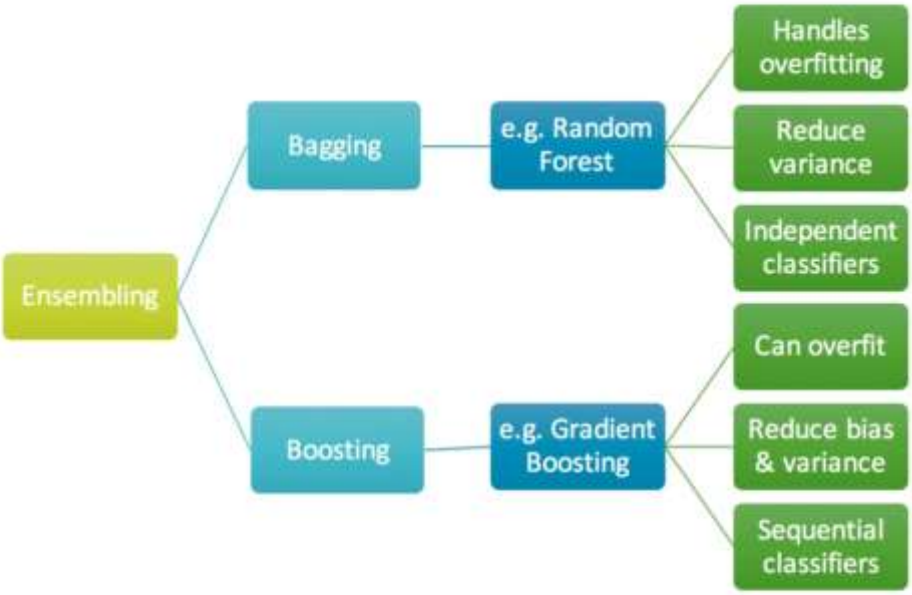

Fig1: Ensemble Learning Techniques (Grover, 2017)

Ensemble learning techniques, such as bagging and boosting, as shown in Fig 1 above, enhance prediction accuracy by combining multiple models. As Random Forest exemplifies, bagging builds independent models from random data subsamples and averages their predictions to reduce variance. Boosting constructs models sequentially, with each model correcting errors of its predecessors, leading to rapid accuracy improvement, as seen in Gradient Boosting. Gradient Boosting Trees (GBTs) are particularly significant in this context, offering a powerful and flexible method that handles complex datasets and improves prediction performance (H. R Sanabila; Wisnu Jatmiko, 2018). Both methods, while distinct in approach, leverage the strengths of multiple predictors to mitigate variance and bias.

4. Boosting Basics

As a fundamental building component of GBT, decision trees function as a base learner. They recursively divide the feature space using the best feasible splitting criteria until the termination requirements are satisfied. While decision trees have various benefits, including interpretability and simplicity of implementation, they are prone to overfitting, instability, and sensitivity to noisy input. Ensemble approaches have emerged as potential options to address these concerns. Boosting algorithms have received much attention because they combine numerous weak learners to create powerful prediction models. Boosting is a supervised learning method that improves the prediction ability of weak learners by repeatedly modifying their weights based on previous performance. It does this by concentrating on cases erroneously categorized in the last round and prioritizing those samples. Consequently, succeeding rounds focus on refining predictions for difficult-to-learn cases, eventually improving overall performance.

AdaBoost (Adaptive Boosting) was among the first and most effective boosting strategies. Its primary goal was to teach a series of poor learners, each of whom was accountable for correcting the flaws of its predecessor. Despite its impressive accomplishments, AdaBoost had certain intrinsic flaws, most notably its sensitivity to irrelevant characteristics and the absence of continuous output support. These considerations laid the groundwork for creating Gradient Boosting (Haohua Wan, 2017).

5. Gradient Boosting: From Theory to Practice

Gradient Boosting expands on the concepts of AdaBoost by introducing optimization techniques, allowing it to address both regression and classification issues. At its core, it incrementally fits decision trees to the residual errors created by the preceding tree in the series. This iterative process continues until a predetermined stopping threshold is met, resulting in efficient convergence and excellent predicted accuracy.

Considering a labeled dataset $D = \{(x_1, y_1),.., (x_n, y_n)\}$ with n observations. Here, xi represents the input vector of dimension d, and yi signifies the target variable. We want to discover a mapping $f(x_i)$ that correctly approximates the link between inputs and targets. Assuming a regression model, the loss function $L(y, f(x))$ measures the difference between the observed goal y and the expected output $f(x)$. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are common loss functions. Cross-entropy loss and logarithmic loss are both often used in classification issues.

Figure 2 shows one of the examples of the proposed model, as indicated in the study conducted by Deng et al. (2019), which involves four key phases utilizing the Gradient Boosting Decision Tree (GBDT) algorithm. First, an automatic data collection program gathers financial data, including company performance metrics and economic indicators, from various financial data sources and databases, covering 30-day, 60-day, and 90-day time windows. In the model training phase, the GBDT algorithm, optimized by Differential Evolution (DE), is used to train the forecasting model. This trained GBDT–DE model then analyzes the financial data to predict future market trends and economic outcomes. Finally, the model's performance is evaluated based on accuracy and efficiency, with an analysis of the importance of each indicator. This approach can be applied as it leverages GBDT's powerful boosting capabilities to enhance the accuracy and reliability of financial forecasting, providing a robust solution for market analysis and decision-making.

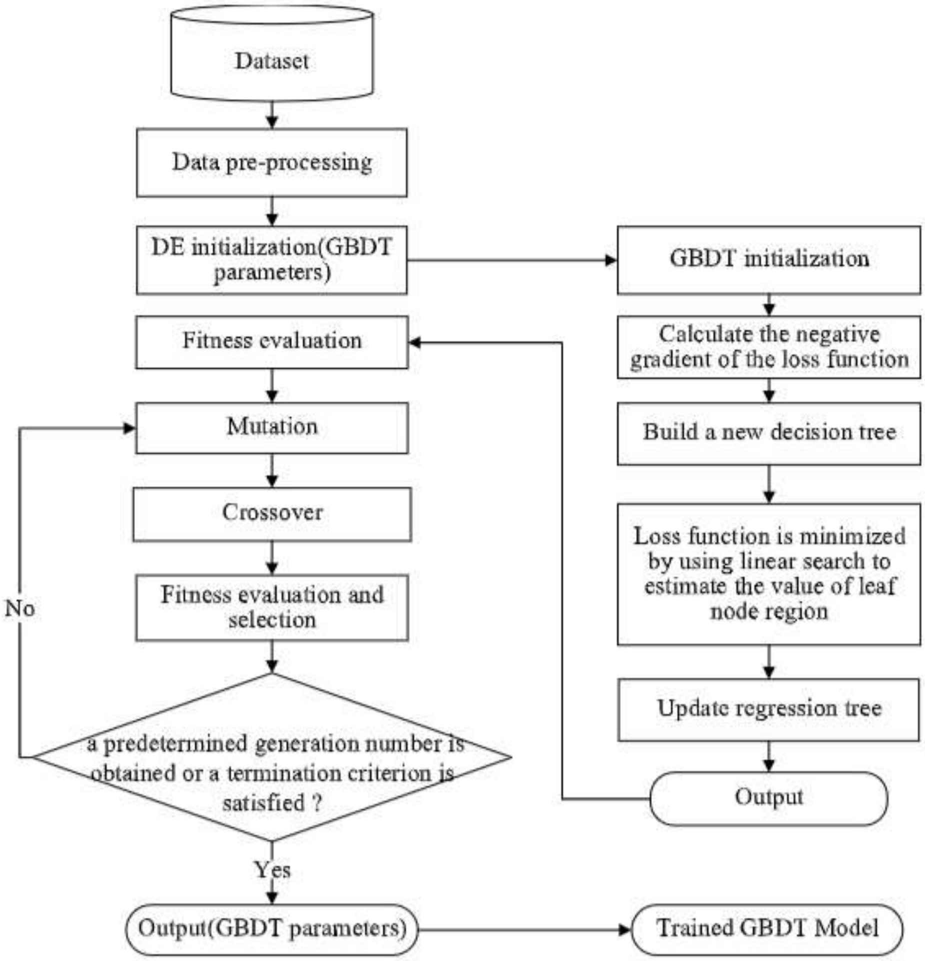
Dataset
Data pre-processing
DE initialization(GBDT parameters)
Fitness evaluation
Mutation
Crossover
Fitness evaluation and selection
a predetermined generation number is obtained or a termination criterion is satisfied ?
No
Yes
Output(GBDT parameters)
GBDT initialization
Calculate the negative gradient of the loss function
Build a new decision tree
Loss function is minimized by using linear search to estimate the value of leaf node region
Update regression tree
Output
Trained GBDT Model

Fig 2: Operational View of Proposed GBDT Model (Deng et al., 2019)

6. Gradient Boosting Trees for Financial Forecasting

In financial forecasting, robust and accurate predictions are paramount. The flowchart in Figure 3 illustrates how gradient-boosting trees can be applied effectively in this domain. Using a comprehensive financial dataset (Dataset D), bootstrap sampling creates multiple subsets, ensuring variability and robustness in the training process. Each subset trains separate decision trees, capturing different financial data patterns. The predictions from these trees are then sequentially refined through the gradient boosting process, where each successive tree focuses on correcting the errors of its predecessors. This method enhances prediction accuracy by minimizing errors and capturing complex patterns in financial markets.

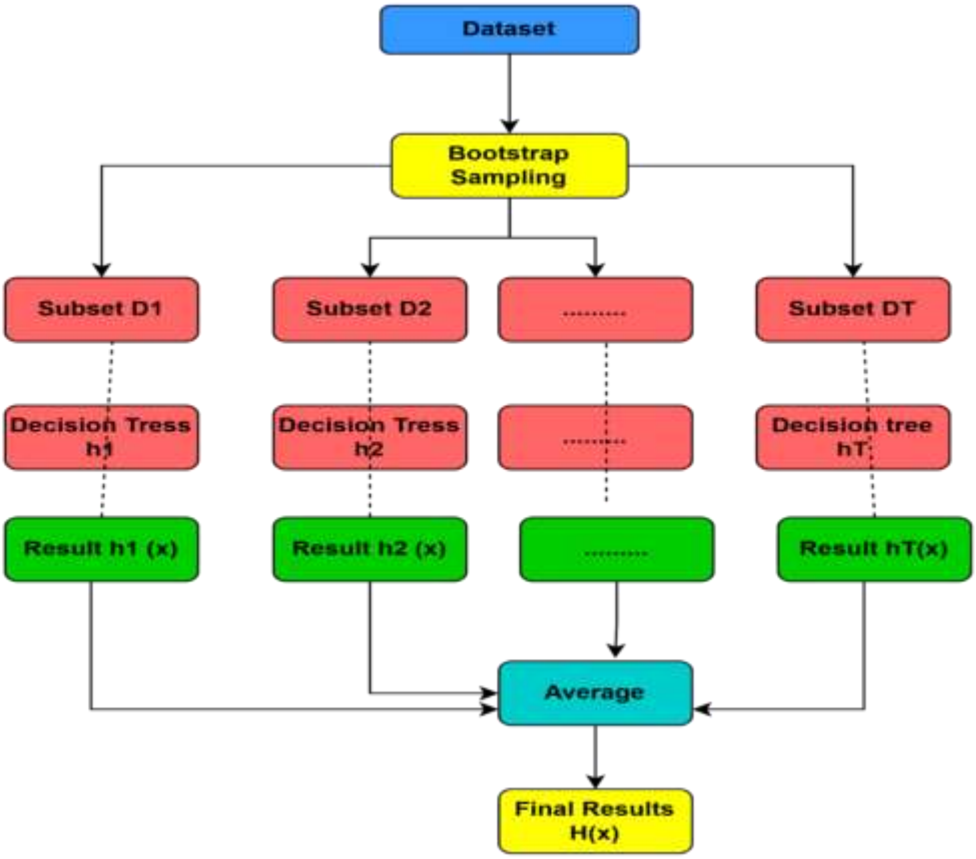

Fig 3: Ensemble Learning using Bootstrap Sampling and Decision Trees

7. Advanced Techniques in XGBM and LightGBM

Regularization techniques in XGBM, such as L1 and L2 regularization, dropout, and early stopping, prevent overfitting and enhance the model's generalization to unseen financial data. Additionally, XGBM's parallel processing capabilities accelerate the training process, enabling timely financial forecasts. On the other hand, LightGBM employs leaf-wise growth and advanced techniques like Gradient-based one-sided sampling (GOSS) and Exclusive Feature Bundling (EFB) to handle large financial datasets efficiently. GOSS ensures that significant under-trained instances are included in the training, while EFB reduces dimensionality without significant information loss(Anghel et al., 2019). These techniques make LightGBM faster and more efficient, which is crucial for processing vast financial data and making accurate, timely predictions.

8. Convergence Properties

As the number of trees increases, the ensemble gradually approaches the ideal solution. Under specific assumptions, the convergence characteristics of GBT may be theoretically investigated. One such assumption is the presence of a finite approximation error limit E, which implies that every good model will ultimately attain acceptable performance. Another crucial assumption is the unbiasedness property, which states that the predicted error of the next tree is zero when conditioned on the present tree's error (Haohua Wan, 2017).

9. Advantages of GBTs in Financial Forecasting

Capturing Complex Relationships: Financially relevant data often display intricate and non-linear relationships between variables shaped by myriad external factors. Traditional linear models struggle to grasp these complexities fully, limiting their suitability for forecasting tasks where a nuanced understanding of relationships is paramount. GBTs, however, offer a more sophisticated perspective by effectively modeling these intricate dependencies (Körner et al., 2018).

Non-linearity. Financial time series data typically contain non-linear trends and seasonality patterns, challenging conventional linear models. GBTs address this issue by iteratively adding trees to the ensemble, each focused on reducing the residual errors generated by its predecessor (Oland, Anders, et al., 2017). This hierarchical structure enables GBTs to approximate complex relationships within the data gradually.

Interactions. In finance, interactions between variables are prevalent, influencing the outcome of interest. Linear models cannot account for these multiplicative effects directly. However, GBTs can implicitly capture interaction terms by combining multiple decision trees.

External factors. Economic indicators, news articles, social media sentiments, and other exogenous factors significantly affect financial outcomes. GBTs accommodate these influences by considering all available information during training, allowing for more accurate forecasts.

Robustness to Overfitting. Overfitting occurs when a model excessively focuses on the noise in the training data, compromising its ability to generalize to new situations. GBTs naturally combat overfitting through their iterative design. Each tree added to the ensemble attempts to minimize the remaining error after accounting for the contributions of earlier trees. Techniques like early stopping and regularization strengthen GBTs against overfitting, ensuring stable and reliable predictions even when dealing with large datasets (Agapitos et al., 2017).

Interpretability. Transparency is essential in finance since stakeholders require clear justification for investment strategies and risk management decisions. Although less interpretable than linear models, GBTs offer valuable insights through the importance of features. These rankings reveal which variables contribute most significantly to the model's predictions, helping analysts understand the underlying drivers shaping financial outcomes.

Scalability. Handling massive datasets is becoming increasingly important in finance due to abundant data sources and the need for real-time analytics. Parallel processing frameworks and computing power advances make applying GBTs to large-scale financial datasets feasible. This scalability empowers organizations to leverage the full potential of their data resources, unlocking new opportunities for innovation and growth.

10. Limitations of GBTs in Financial Forecasting

Gradient Boosting Trees (GBTs) are highly advantageous for financial forecasting because they can effectively capture complex relationships and handle non-linearities in data. However, they come with several notable limitations. Firstly, training GBTs can be computationally intensive and time-consuming, particularly with large datasets and numerous features requiring substantial processing power. Secondly, GBTs are sensitive to hyperparameter tuning, necessitating extensive experimentation to optimize parameters like learning rate and tree depth. Moreover, despite mechanisms to prevent overfitting, GBTs can still succumb to overfitting if not properly regularized or if stopping criteria are not well-defined, especially in complex models. Additionally, their interpretability is often lower than that of simpler models like linear regression, posing challenges in explaining predictions to stakeholders. GBTs also demand high data quality, as noise and irrelevant features can significantly impact performance, necessitating thorough preprocessing. While advances in parallel processing enhance scalability, applying GBTs to large datasets, such as real-time financial data, remains challenging and requires robust infrastructure (Sakata et al., 2018). Lastly, the complexity of setting up and implementing GBTs and the need for specialized knowledge presents a significant learning curve for practitioners new to machine learning techniques.

Table 1 provides a comprehensive overview of GBTs' advantages and limitations in financial forecasting, highlighting their strengths in capturing data complexities and the practical considerations required for their effective implementation.

Table 1: Advantages and Limitations of Gradient Boosting Trees (GBTs)

Feature	Advantages	Limitations
Complex Relationships	Models intricate and non-linear dependencies in financial data.	Training can be computationally expensive and time-consuming.
Non-linearity Handling	Addresses non-linear trends and seasonality patterns.	Requires careful tuning of hyperparameters.
Interaction Terms Capture	Captures interaction terms by combining multiple decision trees.	Risk of overfitting if not properly regularized.
External Factors	Considers diverse data sources like economic indicators and social media sentiments.	Less interpretable than simpler models.
Robustness to Overfitting	It uses techniques like early stopping to prevent overfitting.	Highly sensitive to data quality.
Feature Importance	Provides insights through feature importance rankings.	Handling extensive datasets can be challenging.
Scalability with Large Datasets	Parallel processing enables handling massive datasets.	Complex to set up and implement.
Parameter Sensitivity	Tuning parameters allows for fine-tuning the model's performance.	Extensive experimentation is needed to find optimal settings.
Implementation Complexity	Allows for sophisticated and robust models that can handle complex data and tasks.	Requires sophisticated tools and a steep learning curve for practitioners.

11. Application Areas of GBTs in Financial Forecasting

Stock Market Prediction

Historical Performance. Numerous studies have compared the predictive prowess of GBTs versus traditional statistical models in stock market forecasting. Research indicates that GBTs consistently outperform Autoregressive Integrated Moving Average (ARIMA) models in capturing the volatility and non-linear dynamics characteristic of financial markets.

Feature Selection. Incorporating relevant financial indicators and macroeconomic variables into the GBT model can improve forecasting results. Features like moving averages, momentum indices, and technical indicators help identify emerging trends and patterns, contributing to more accurate predictions.

Real-time Processing. High-frequency trading demands rapid response times, necessitating real-time data processing. Thanks to their parallelizable architecture and efficient training algorithms, GBTs can be deployed in real-time environments (Jiao & Jakubowicz, 2017).

Credit Risk Assessment

Improved Scoring Models. Banks and financial institutions rely on credit risk assessment models to evaluate borrower creditworthiness and manage loan portfolios effectively. To build accurate credit scoring models, GBTs can analyze historical data on loan repayment behavior, income levels, employment history, and other financial indicators.

Adaptivity. GBTs can adapt to borrower profiles and changes in economic conditions, ensuring that credit risk scores remain up-to-date and reflect current circumstances.

Fraud Detection. Identifying fraudulent activities in credit applications requires sophisticated analytical tools to distinguish genuine cases from malicious ones. GBTs can effectively handle imbalanced datasets and adapt to evolving fraud tactics, bolstering the security measures employed by financial institutions (Chang et al., 2018).

Fraud Detection

Anomaly Detection: Detecting anomalous patterns in transactional data is crucial for preventing financial losses due to fraudulent activities. GBTs excel in this domain by identifying subtle deviations from standard transaction patterns and flagging potentially fraudulent transactions in real-time.

Evolutionary Threats. As fraudsters adopt new tactics to circumvent existing detection systems, GBTs can be retrained using updated data to maintain efficacy. This adaptability ensures that financial institutions stay ahead of evolving threats and protect their assets.

Compliance. Meeting regulatory requirements related to anti-money laundering and knowing customer regulations necessitates rigorous monitoring and reporting. GBTs can streamline compliance efforts by automating identifying and escalating suspicious transactions, freeing up valuable resources for other mission-critical tasks (Mishra & Ghorpade, 2018).

12. Volatility Estimation

Estimating volatility is a crucial component of managing financial risk and pricing options. The process entails quantifying the variability or dispersion of the financial instrument's returns throughout a designated timeframe. Accurate volatility evaluation allows investors to assess the risks linked to their assets and make educated choices about portfolio allocation and hedging measures. Historically, the volatility estimate depended on the use of Autoregressive Conditional Heteroskedasticity (ARCH) models and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models. These statistical models rely on historical data and assume that the underlying asset's volatility follows a specific time series pattern. However, they may have difficulty accurately capturing complex patterns and interactions among several elements that influence volatility (Guo et al., 2018).

Lately, academics have paid considerable attention to using machine learning methods, namely Gradient Boosting Trees (GBT), to estimate volatility. Gradient Boosting Trees (GBT) is a flexible machine learning approach often used for regression problems. This is due to its ability to identify non-linear relationships between characteristics and the target variable iteratively, including weak decision trees. The "GBT-ARCH" technique is formed by combining GBT and ARCH models. GBT is used to forecast the average value of the returns, whereas ARCH models estimate the volatility of the residual errors. The combined estimates result in improved accuracy as compared to independent ARCH models. Several other researchers proposed a hybrid model called "GBTLSTM." This novel approach combines Gradient Boosting Trees (GBT) and Long Short-Term Memory (LSTM) networks to predict volatility accurately. LSTMs are advanced, recurrent neural network architectures that excel at capturing complex temporal relationships seen in sequential data. GBTLSTM combines GBT and LSTM components to detect complex relationships between input characteristics and volatility. It achieves outstanding results on several benchmark datasets compared to other state-of-the-art approaches.

Machine learning methods, such as Gradient Boosting Trees (GBT) and Long Short-Term Memory (LSTM), provide significant advantages compared to traditional statistical models regarding estimating volatility. They do not make assumptions about the distribution of the underlying data and may effectively understand complex correlations between qualities and goals utilizing large amounts of information. Furthermore, they enable the processing of many inputs, such as macroeconomic information and market mood indices, improving the ability to analyze volatility. However, specific obstacles are associated with applying machine learning models for volatility estimates. One major challenge is ensuring enough accessible training data, especially for rare assets or events that do not happen often and have limited historical records. Interpretability is a significant challenge for opaque models such as deep learning frameworks. These models do not provide insight into how they function, making it difficult to understand the factors contributing to volatility (Kumar & Patil, 2018).

Despite these challenges, research on machine learning methods, particularly GBT and its variations, for estimating volatility shows significant advancements. As computing resources increase and data quantities grow, we expect improvements and refinements in these models. This will ultimately result in more robust and efficient risk management applications in the financial industry.

13. Macroeconomic Forecasting

Macroeconomic forecasting involves predicting national-level economic variables, such as inflation rates, interest rates, and Gross Domestic Product (GDP). Several studies have explored the use of Gradient Boosting Trees (GBT) in macroeconomic forecasting, demonstrating its capacity to outperform traditional econometric models (Döpke et al., 2017). Researchers have used Gradient Boosting Trees (GBT) to forecast the quarterly growth rates of the US Real GDP. Instead of utilizing past data, they have employed leading indicators. GBT outperformed ARIMA and Seasonal ARIMA models in predicting quarterly US actual GDP growth rates. Monthly Chinese Industrial Production Index Forecast: This is another example demonstrating the efficacy of GBT in predicting the Monthly Chinese Industrial Production Index. GBT yielded superior results to ARIMA and Exponential Smoothing State Space Models (Li & Zhang, 2018).

These results highlight the capacity of GBT in macroeconomic forecasting. Conventional econometric models often depend only on past data and predetermined mathematical structures. On the other hand, GBT can adjust and learn from different inputs, making it a more versatile option for predicting economic factors. GBT can include several sets of pertinent economic statistics, such as consumer price index, unemployment rate, housing starts, and factory orders, to provide forecasts. Furthermore, it can include non-linear connections and changing patterns that may not be accounted for by fixed econometric models. Moreover, GBT can effectively handle high-dimensional feature spaces and address the challenge of missing data points often encountered in macroeconomic forecasting. The fact that it can tackle these issues makes it a potential option for dealing with the intricacies present in macroeconomic systems. Nevertheless, it is crucial to recognize some constraints of Gradient Boosting Trees (GBT) in macroeconomic forecasting. Interpreting the model's results may be difficult because of its opaque nature. Furthermore, it is crucial to thoroughly assess the model's capacity to apply to various situations and areas before implementing it on a large scale. The development of transparent iterations of GBT should be prioritized, enabling users to comprehend the reasoning behind the model's predictions (Touzani et al., 2018). Additionally, it is worth considering ensemble approaches that combine Gradient Boosting Trees (GBT) with other forecasting techniques to enhance the forecasts' overall accuracy and dependability. Incorporating machine learning methods, namely GBT, into macroeconomic forecasting has excellent potential. It may provide a valuable understanding of the changing dynamics of global economies and facilitate proactive policy responses.

14. Improving Performance in Financial Forecasting

Financial forecasting is a complex and dynamic field where precise prediction models are crucial. Gradient Boosted Trees (GBT) have emerged as a powerful machine-learning technique in this domain due to their ability to handle complex data structures and interactions. However, the performance of GBT models can be significantly enhanced through various techniques, including feature selection, ensemble methods, and hyperparameter tuning. This article explores these techniques and their application in financial forecasting, mainly focusing on stock price prediction.

Feature Selection Techniques. Feature selection is a critical step in building robust machine-learning models. It involves identifying and selecting the most relevant features from a dataset while reducing dimensionality. This process improves model interpretability and enhances generalization performance by preventing overfitting (Bhalaji et al., 2018). Several methods have been proposed for feature selection in GBT models used for financial forecasting. One practical approach combines mutual information maximization and recursive feature elimination. Mutual information maximization measures the dependency between variables, helping identify features that provide the most information about the target variable. Recursive feature elimination, on the other hand, iteratively removes the most minor essential features based on model performance until the optimal subset is obtained. This combined approach has enhanced predictive accuracy in the context of stock price prediction. Selecting a subset of highly informative features makes the model more efficient and less prone to overfitting, leading to more reliable forecasts. Additionally, feature selection helps in reducing computational complexity, making the training process faster and more efficient.

15. Ensemble Methods

Ensemble methods have proven to be highly effective in improving the performance of machine learning models. These methods combine multiple base models to produce a more robust and accurate prediction than any single model. The fundamental idea behind ensemble methods is that different models can capture various aspects of the data, and their combination can leverage these diverse insights. In financial forecasting, several ensemble methods have been explored with GBT. One notable approach is stacking, which involves training multiple GBT models on the same dataset and then using their predictions as inputs to a final meta-model. This technique allows the strengths of various models to be combined, resulting in improved prediction accuracy. For example, in stock price prediction, a stacking framework consisting of multiple GBT models has demonstrated superior performance to individual GBT models. By aggregating the predictions from various models, the ensemble method can mitigate the weaknesses of any single model, leading to more reliable and accurate forecasts (H. R Sanabila; Wisnu Jatmiko, 2018). Another ensemble method that has been applied in financial forecasting is boosting. Boosting involves training models sequentially, where each new model attempts to correct the errors of the previous ones. This iterative process results in an intense final model that is highly accurate. In the context of GBT, boosting is inherently built into the algorithm, as it constructs an ensemble of decision trees where each tree is trained to correct the errors of the preceding ones.

16. Hyperparameter Tuning

Hyperparameters play a crucial role in the performance of machine learning models. Unlike model parameters learned during training, hyperparameters are set before training and must be carefully tuned to optimize model accuracy (Andreea, et al., 2018). Several strategies have been proposed for hyperparameter tuning in GBT models. The most common methods include grid search, random search, and Bayesian optimization. Grid search involves systematically searching through a predefined set of hyperparameters, evaluating model performance for each combination. While this method is exhaustive, it can be computationally expensive and time-consuming.

Random search, in contrast, randomly samples hyperparameter combinations from a specified distribution. This approach is often more efficient than grid search, as it does not evaluate all possible combinations but can still find a near-optimal set of hyperparameters. Bayesian optimization is a more sophisticated method that uses probabilistic models to model the relationship between hyperparameters and model performance. By iteratively updating this model based on previous evaluations, Bayesian optimization can efficiently explore the hyperparameter space and identify the optimal configuration. This approach has significantly improved GBT models' performance in financial forecasting (Andreea et al., 2018). In stock price prediction, employing Bayesian optimization for hyperparameter tuning has resulted in superior performance compared to manually selected hyperparameters. This automated approach saves time and computational resources and ensures the model operates at its best possible configuration, leading to more accurate forecasts.

Table 2: Various GBDT Applications and Accuracies

Ref.	**Authors**	**Model**	**Concept**	**Accuracy / F1 Score**
(Xia et al., 2017)	Xia, Y., Liu, C., Li, Y., & Liu, N.	GBDT **XGBoost-TPE**	Predicting Credit scoring for banks to properly guide decisions profitably on granting loans.	GBDT - 86.14 **XGBoost-TPE - 87.92**
(Zhang & Meng, 2018)	Zhang, T., & Meng, S.	GBDT + Logistic Regression	Credit evaluation	GBDT- 83.5% GBDT + LR - 85.8%
(Li, 2018)	Li, Z.	GBDT-SVM	Credit Risk assessment	96.95%
(Wyrobek, 2018)	Joanna, W. (2018).	GB	Predicting Bankruptcy	93.8

Table 2 above provides a glimpse into various applications of Gradient Boosting (GBDT) and its accuracy in financial risk assessment and credit scoring techniques. Xia et al. demonstrated the effectiveness of XGBoost-TPE, achieving high accuracy rates of 86.14% and 87.92% in predicting credit scores, emphasizing its utility for banks in making informed loan decisions. Zhang and Meng integrated GBDT with logistic regression to enhance credit evaluation, achieving accuracies of 83.5% and 85.8%, respectively, showcasing a hybrid approach's effectiveness in improving predictive performance. Li utilized GBDT-SVM integration to achieve a notably high accuracy of 96.95% in credit risk assessment, underscoring the robustness of ensemble methods in complex financial tasks. Lastly, Joanna's work highlighted GBDT's efficacy with a 93.8% accuracy in predicting bankruptcy, illustrating its versatility in diverse financial prediction scenarios. These studies collectively underscore GBDT's role as a powerful tool in financial analytics, offering both high accuracy and practical applicability in decision-making processes within banking and finance.

Conclusion

To summarize, this research thoroughly examined using gradient-boosted trees (GBT) in financial forecasting. We explored the benefits of Gradient Boosting Trees (GBT) compared to other widely used machine learning algorithms. Additionally, we showcased many applications of GBT in financial forecasting, such as predicting stock prices, estimating volatility, and forecasting macroeconomic trends. In addition, we examined several approaches used to improve the efficiency of GBT in financial prediction, including feature selection strategies, ensemble methods, and hyperparameter tweaking. Our study emphasizes the increasing significance of machine learning models such as GBT in financial forecasting. It also indicates areas for further investigation, namely in overcoming obstacles with the comprehensibility and interpretability of these models.

References

[1]. Agapitos, A., Brabazon, A., & O'Neill, M. (2017). Regularised gradient boosting for financial time-series modelling. *Computational Management Science*, *14*(3), 367-391. https://doi.org/10.1007/s10287-017-0280-y

[2]. Anghel et al., A. (2019). Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. *Cornell University*. https://doi.org/10.48550/arXiv.1809.04559

[3]. Bakar, N. A., & Rosbi, S. (2017). Data clustering using autoregressive integrated moving average (ARIMA) model for Islamic country currency: An econometrics method for Islamic financial engineering. *The International Journal of Engineering and Science*, *06*(06), 22-31. https://doi.org/10.9790/1813-0606022231

[4]. Bhalaji, N., Kumar, K. S., & Selvaraj, C. (2018). Empirical study of feature selection methods over classification algorithms. *International Journal of Intelligent Systems Technologies and Applications*, *17*(1/2), 98. https://doi.org/10.1504/ijista.2018.091590

[5]. Chang, Y., Chang, K., & Wu, G. (2018). Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029

[6]. Deng, S., Wang, C., Wang, M., & Sun, Z. (2019). A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market. *Applied Soft Computing*, *83*, 105652. https://doi.org/10.1016/j.asoc.2019.105652

[7]. Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting*, *33*(4), 745-759. https://doi.org/10.1016/j.ijforecast.2017.02.003

[8]. Grover, P. (2017, December 13). Gradient boosting from scratch. *Medium*. https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d

[9]. Guo, T., Bifet, A., & Antulov-Fantulin, N. (2018). Bitcoin volatility forecasting with a glimpse into buy and sell orders. *2018 IEEE International Conference on Data Mining (ICDM)*. https://doi.org/10.1109/icdm.2018.00123

[10]. H. R Sanabila; Wisnu Jatmiko. (2018). Ensemble Learning on Large Scale Financial Imbalanced Data. *2018 International Workshop on Big Data and Information Security (IWBIS)*. https://doi.org/10.1109/IWBIS.2018.8471702

[11]. Haohua Wan. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *courses.grainger.illinois.edu*.

[12]. Ilbeigi, M., Castro-Lacouture, D., & Joukar, A. (2017). Generalized autoregressive conditional Heteroscedasticity model to quantify and forecast uncertainty in the price of asphalt cement. *Journal of Management in Engineering*, *33*(5). https://doi.org/10.1061/(asce)me.1943-5479.0000537

[13]. Jiao, Y., & Jakubowicz, J. (2017). Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks. *2017 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata.2017.8258518

[14]. Kumar, H. P., & Patil, B. S. (2018). Forecasting volatility trend of INR USD currency pair with deep learning LSTM techniques. *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*. https://doi.org/10.1109/csitss.2018.8768767

[15]. Körner et al., P. (2018). Advantages of GBTs in Financial Forecasting. *Technische Universität Dresden*.

[16]. Li, P., & Zhang, J. (2018). A new hybrid method for China's energy supply security forecasting based on ARIMA and XGBoost. *Energies*, *11*(7), 1687. https://doi.org/10.3390/en11071687

[17]. Li, Z. (2018). GBDT-SVM credit risk assessment model and empirical analysis of peer-to-peer borrowers under consideration of audit information. *Open Journal of Business and Management*, *06*(02), 362-372. https://doi.org/10.4236/ojbm.2018.62026

[18]. Mishra, A., & Ghorpade, C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. https://doi.org/10.1109/sceecs.2018.8546939

[19]. Oland, Anders, et al. (2017). Be Careful What You Backpropagate: A Case For Linear Output Activations & Gradient Boosting. *arXiv preprint arXiv:1707.04199*.

[20]. Sakata, R., Ohama, I., & Taniguchi, T. (2018). An extension of gradient boosted decision tree incorporating statistical tests. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw.2018.00139

[21]. Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533-1543. https://doi.org/10.1016/j.enbuild.2017.11.039

[22]. Wyrobek, J. (2018). Predicting bankruptcy at Polish companies: A comparison of selected machine learning and deep learning algorithms. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, (6(978)), 41-60. https://doi.org/10.15678/znuek.2018.0978.0603

[23]. Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225-241. https://doi.org/10.1016/j.eswa.2017.02.017

[24]. Zhang, T., & Meng, S. (2018). Internet financial credit evaluation based on the fusion of GBDT and LR. *Proceedings of the 2018 International Conference on Management, Economics, Education and Social Sciences (MEESS 2018)*. https://doi.org/10.2991/meess-18.2018.17