NTT STOCK PRICE PREDICTION

Presentation by Saurav Raj

Deepcraft

OBJECTIVE & OVERVIEW

Objective : To predict NTT stock prices using various machine learning models to assess their effectiveness

Model Evaluted: Baseline Naïve, Random Forest, LSTM, and an Ensemble Weighted model.

Purpose: A comparative analysis of different machine learning techniques for stock price forecasting. The goal is to determine the best model for accurate stock price predictions, emphasizing both precision and generalizability.

BACKGROUND

Importance of Stock Price Prediction

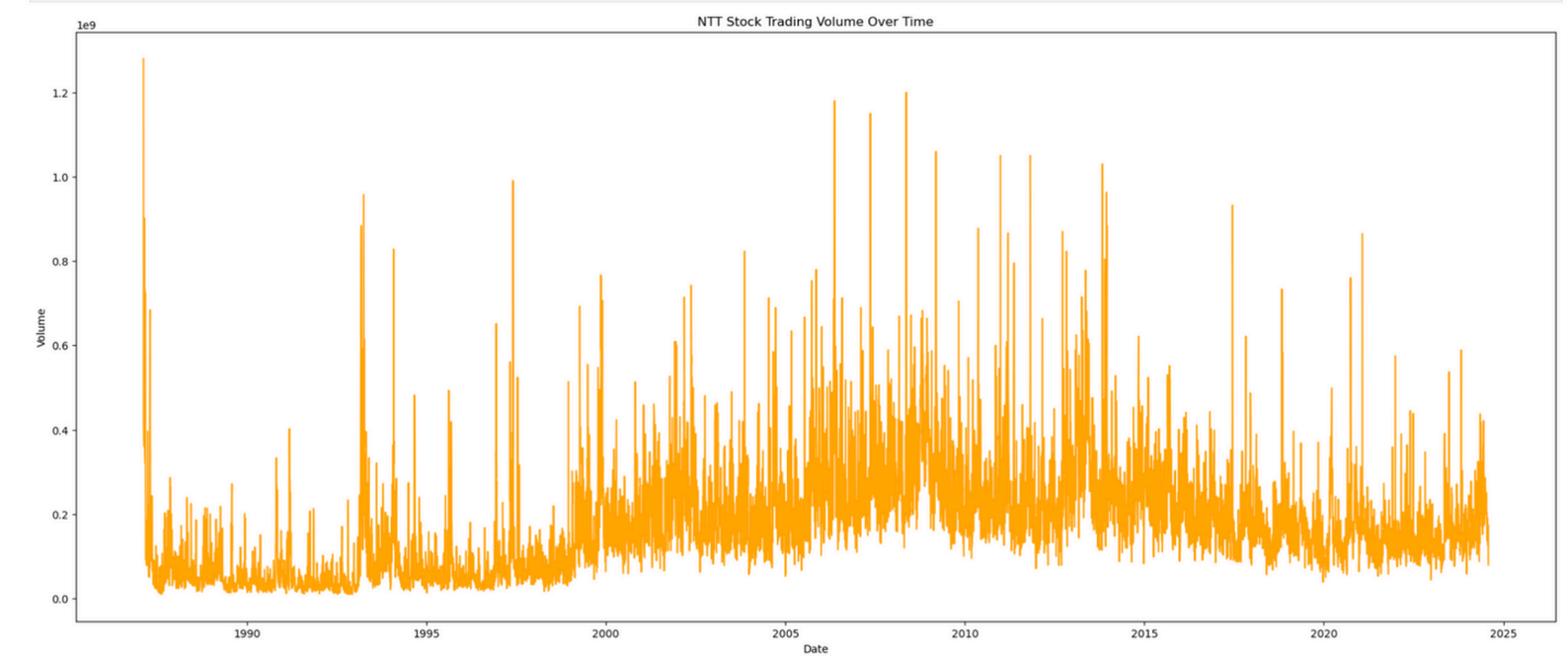
- **Investment Strategy:** Helps investors identify trends and make informed buy or sell decisions to maximize profits
- **Risk Management :** Prediction models help reduce financial risks by forecasting downturns.
- **Market Analysis:** Provides insights into company performance, economic conditions,investor sentiment.

Challenges in Predicting Stock Prices

- **High Volatility:** Stock prices are influenced by numerous unpredictable factors, including market sentiment, economic events, and political decisions
- **Noise in Data:** Financial time-series data is noisy,with random fluctuations that make prediction challenging.
- **Overfitting Risk:** Models may overfit to historical data and perform poorly on unseen data, reducing their practical usefulness.

DATA ANALYSIS

Exploratory Data Analysis (EDA) Results

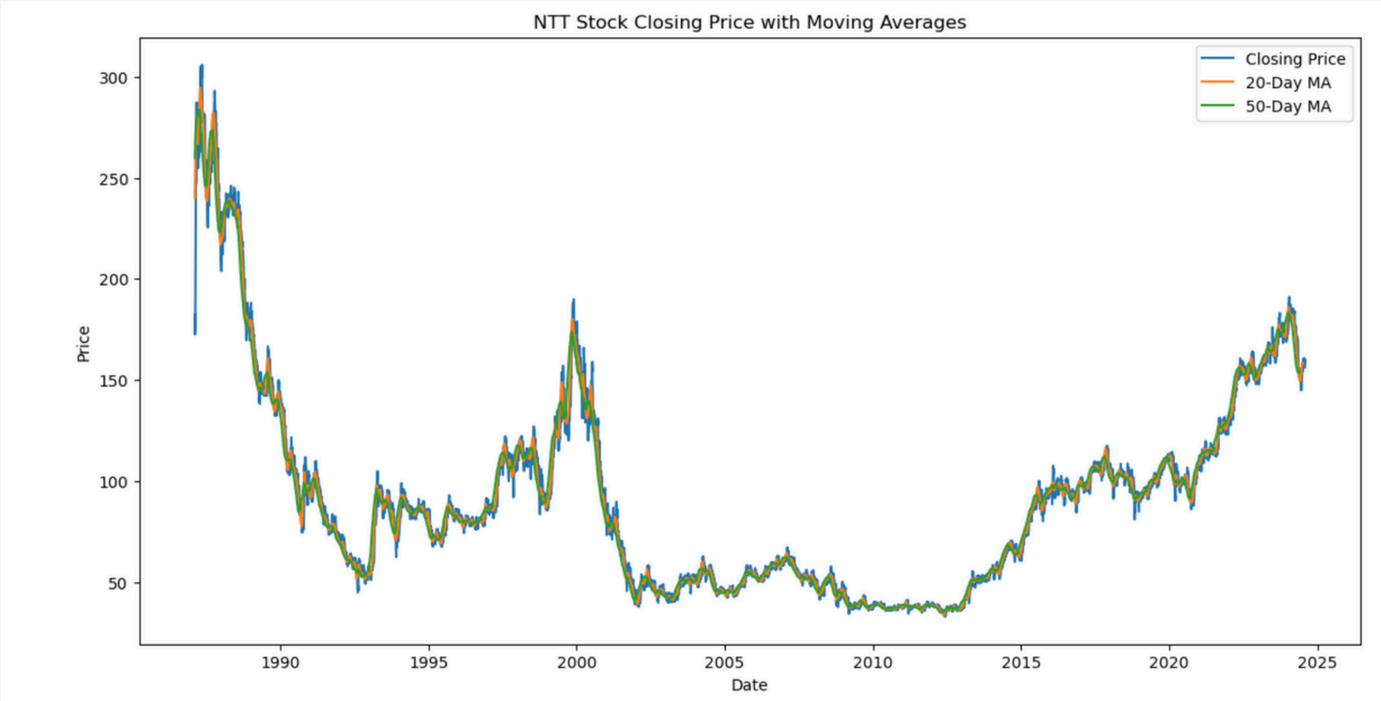


a) Volume Analysis :

Trading volume shows clear seasonality and spikes during certain periods, possibly indicating important market events.

DATA ANALYSIS

Exploratory Data Analysis (EDA) Results



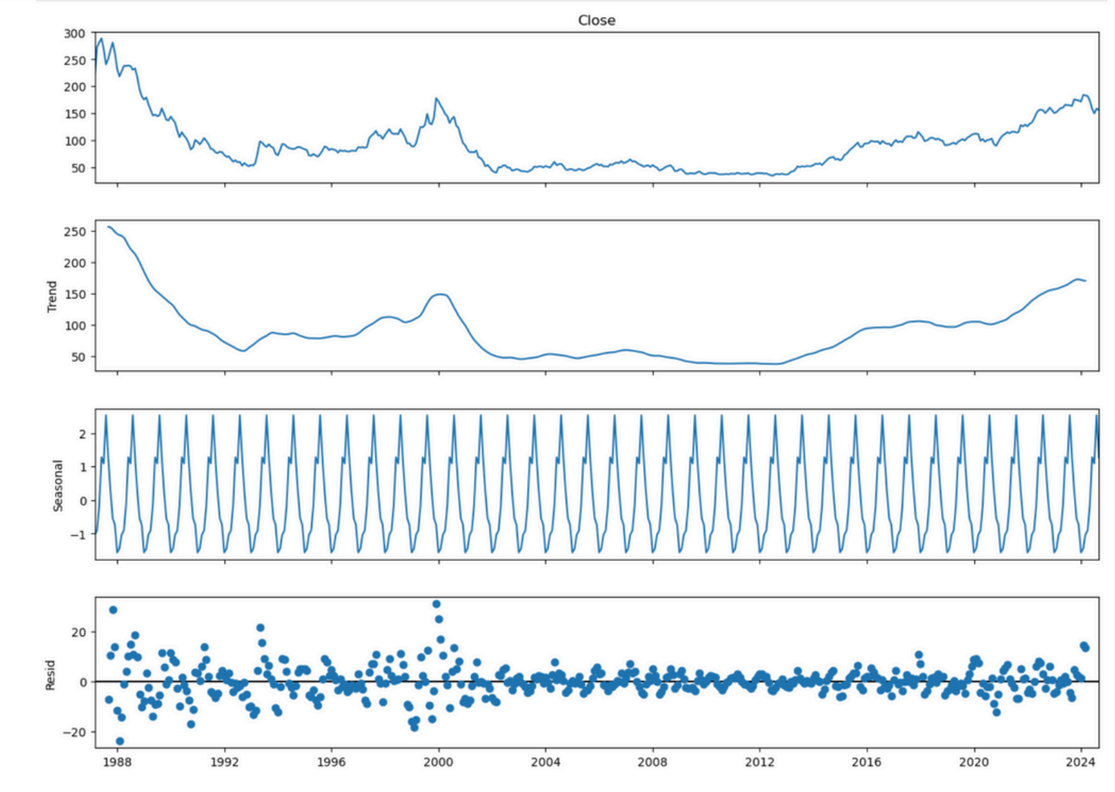
b) Closing Price Trends :

Identified long-term upward and downward trends along with short-term fluctuations

Moving averages (20-day and 50-day) were computed to smooth out noise.

DATA ANALYSIS

Exploratory Data Analysis (EDA) Results



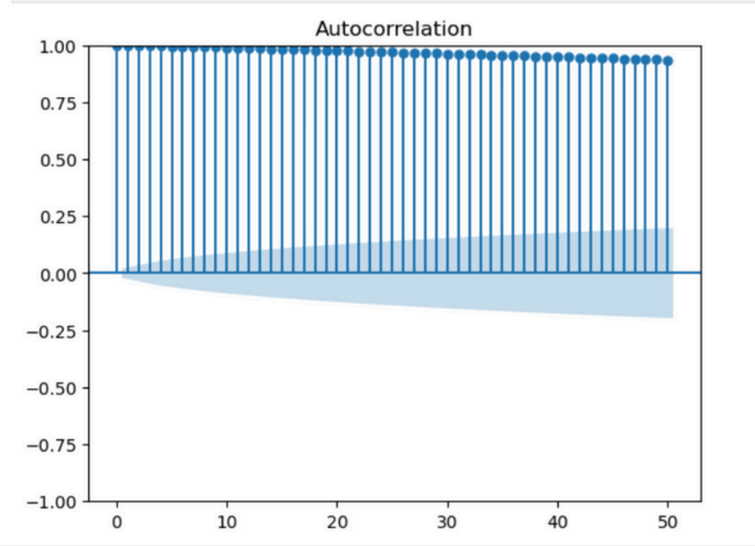
c) Seasonality and Trends :

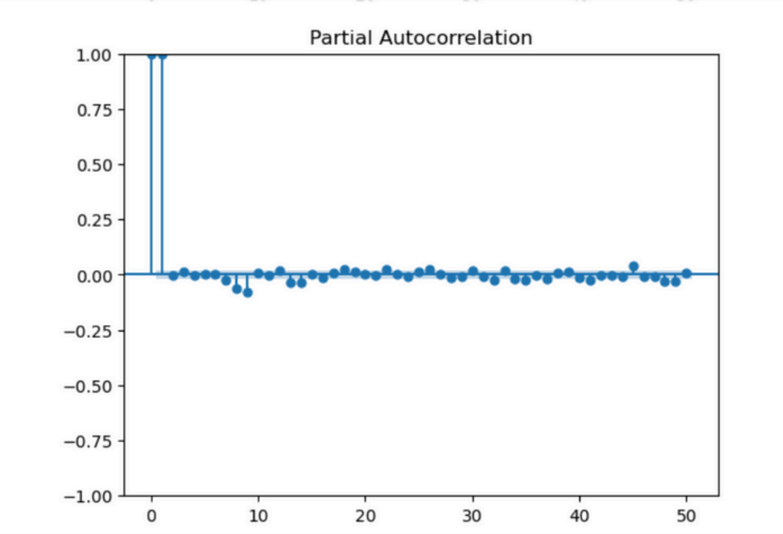
Seasonal decomposition showed significant seasonal patterns in the data.

Autocorrelation plots indicated a high correlation with recent past values, suggesting predictability.

DATA ANALYSIS

Exploratory Data Analysis (EDA) Results



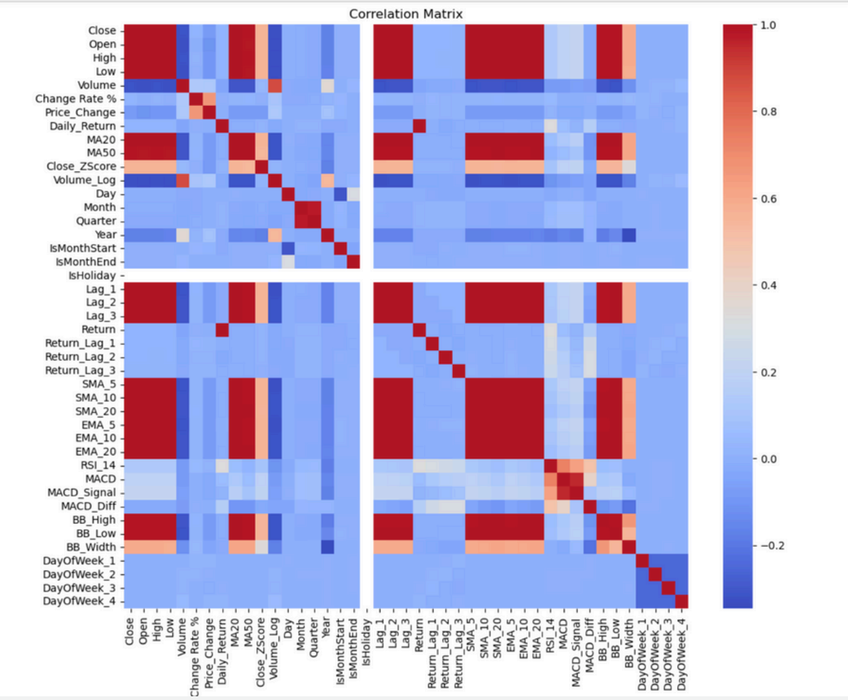


d) Seasonality and Trends :

Autocorrelation plots indicated a high correlation with recent past values, suggesting predictability.

DATA ANALYSIS

Exploratory Data Analysis (EDA) Results

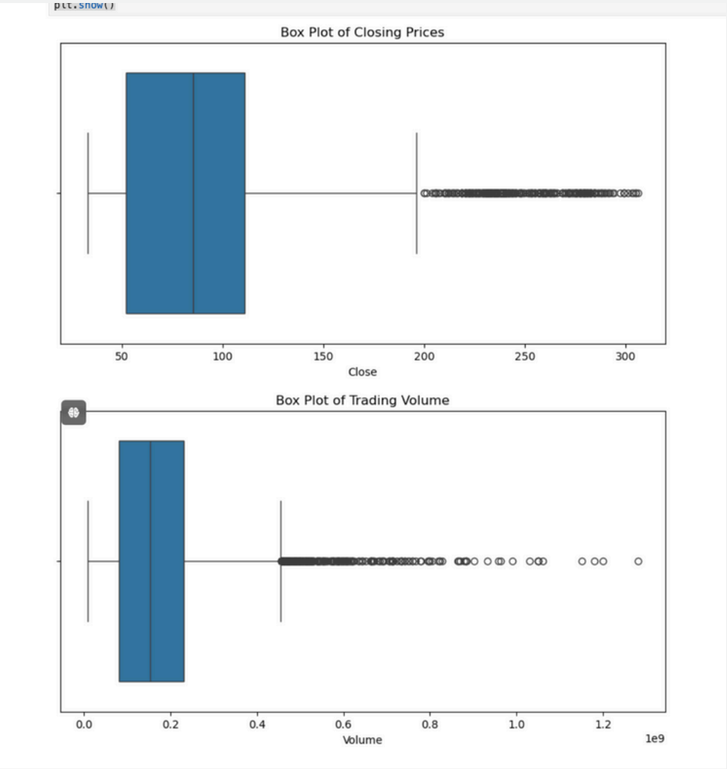


e) Correlation Analysis

Strong correlations were found between features like Open, High, Low, and Close, leading to multicollinearity concerns.

DATA ANALYSIS

Identified Issues



a) High Multicollinearity

Many features like Open, High, Low, and Close had correlations close to 1. This required careful feature selection to avoid redundancy and improve model efficiency.

b) Outliers

Outliers were detected in the closing price and trading volume, indicating potential market anomalies.

DATA ANALYSIS

Identified Issues

c) Noise and Volatility

Significant noise and volatility in daily returns were observed, making prediction more challenging.

Action taken

a) Removed or transformed highly correlated features to reduce multicollinearity.

b) Normalized and scaled the data to handle large variations in volume and price.

c) Used moving averages and technical indicators (e.g., RSI, MACD) to extract more meaningful patterns from noisy data.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used

a) Baseline Naïve Forecast Model :

- Provided a simplistic benchmark for performance comparison. It helped us understand if the more sophisticated models were truly adding value beyond basic historical averaging.
- Baseline Naïve Forecast MAE: 1.0834 (Used as a baseline for assessing advanced models).

b) Random Forest Model ::

- Selected for its robustness to overfitting and its ability to handle a variety of feature types, including highly correlated variables. Random Forest's ensemble nature also helps in improving accuracy by reducing variance.
- By utilizing important features, it managed to learn the relationships between features effectively, outperforming ARIMA and LSTM on the generalization task.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used

b) Random Forest Model :

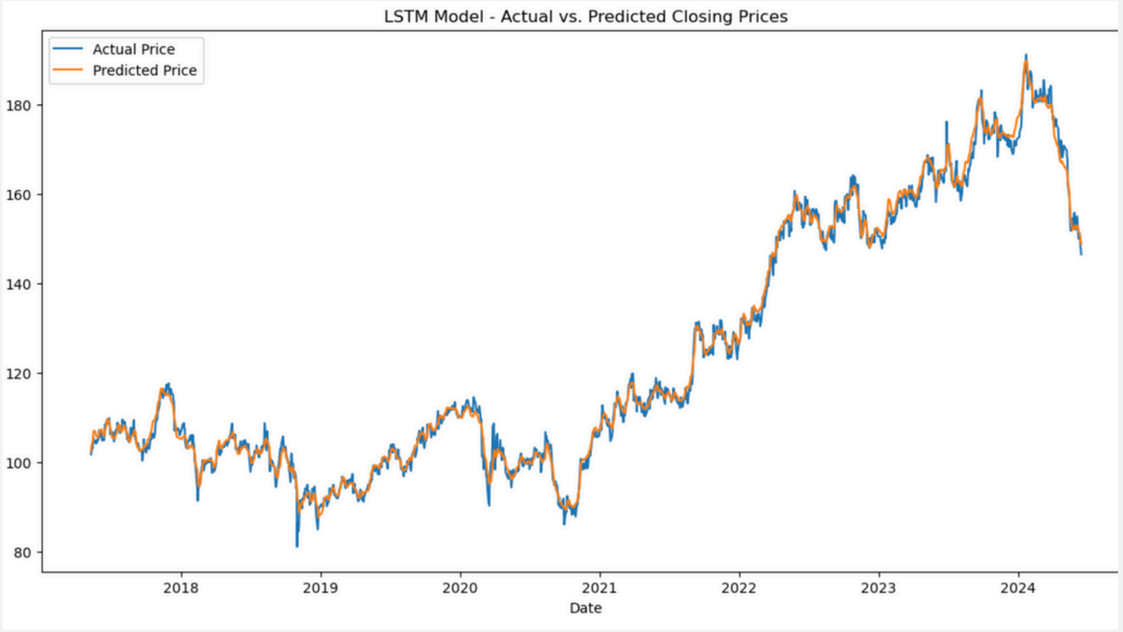
Random Forest: Best hyperparameters

- max_depth: 5
- max_features: 0.3
- min_samples_leaf: 10
- min_samples_split:10
- n_estimators: 50

Random Forest Model MAE: 0.8536 (Indicating solid generalization capability).

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used

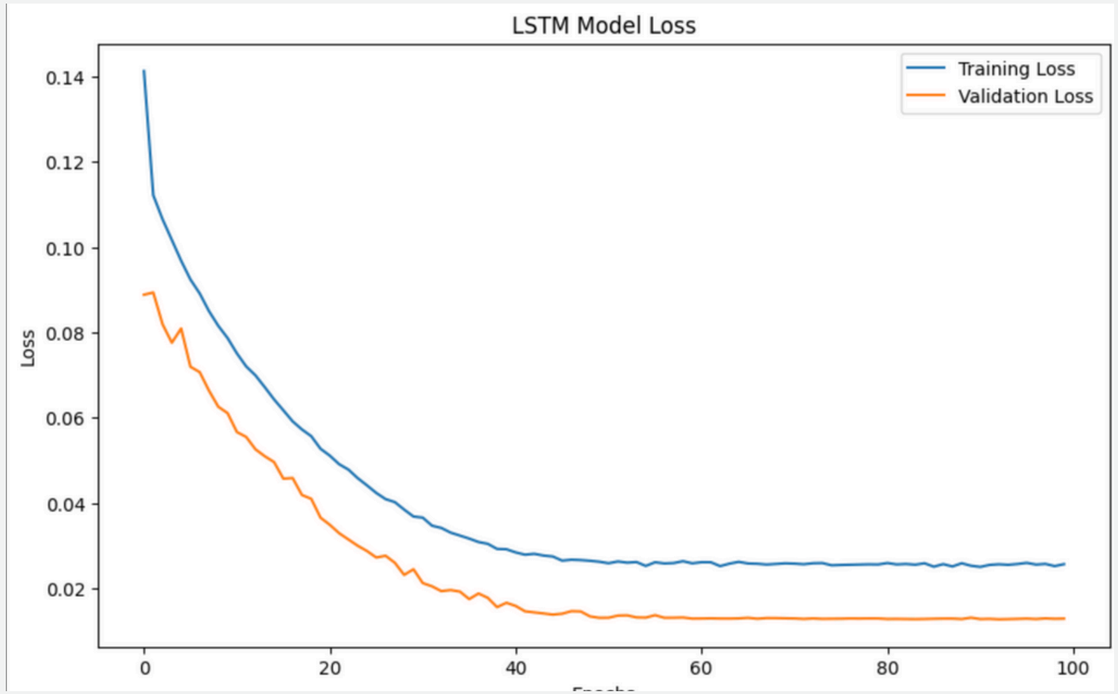


c) LSTM (Long Short-Term Memory) :

- Given that stock prices are sequential in nature, LSTMs are designed to learn temporal dependencies over long sequences. This made LSTM an ideal candidate to model the temporal patterns in stock price movement.
- Despite its potential, LSTM faced difficulties in generalizing well due to high volatility in the data, leading to higher errors compared to Random Forest.
- LSTM Model MAE: 1.3206 (Reflecting that LSTM was good but not the most optimal model compared to Random Forest).

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used

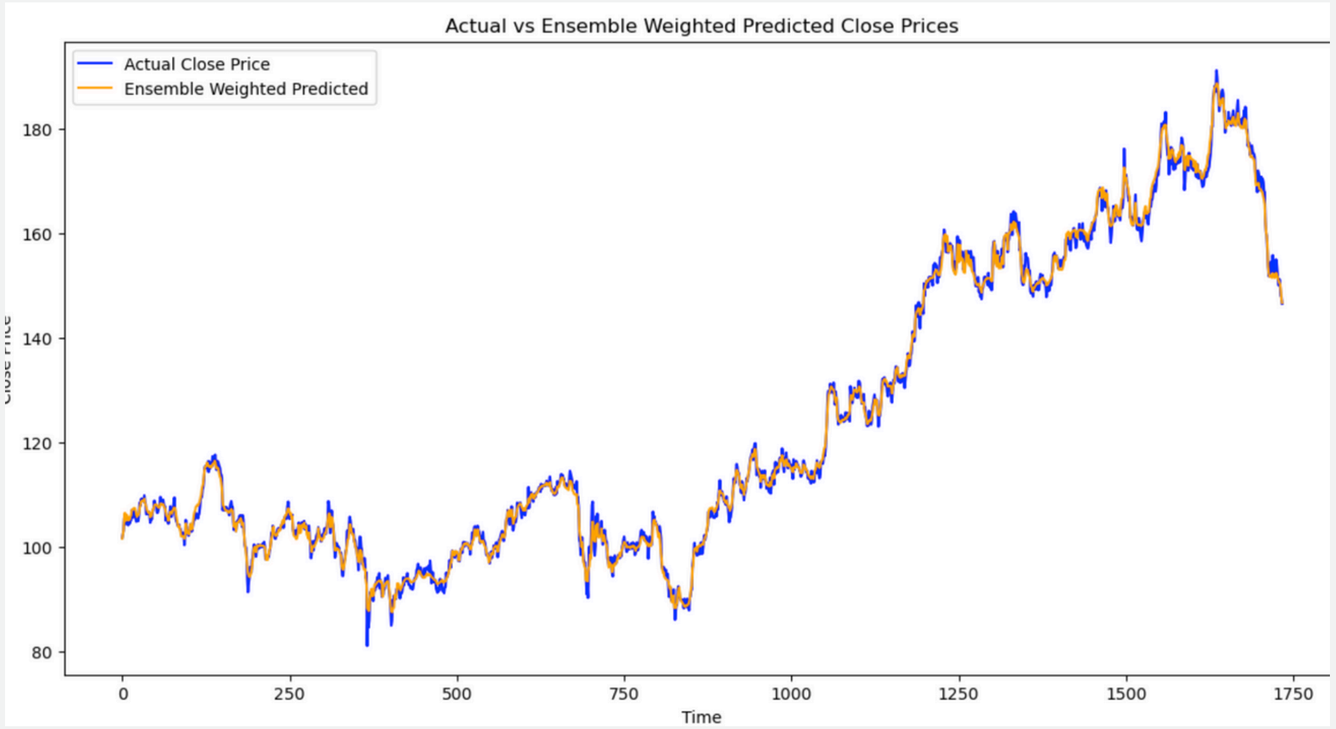


c) LSTM (Long Short-Term Memory) :

- Applied dropout regularization (0.2) to prevent overfitting
- Reduced learning rate to 0.0003 for smoother convergence.
- Early stopping applied with patience of 5 epochs to avoid overtraining
- Hyperparameter tuning ensures that models are well-suited to the data's complexity, allowing them to achieve a balance between bias and variance

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used



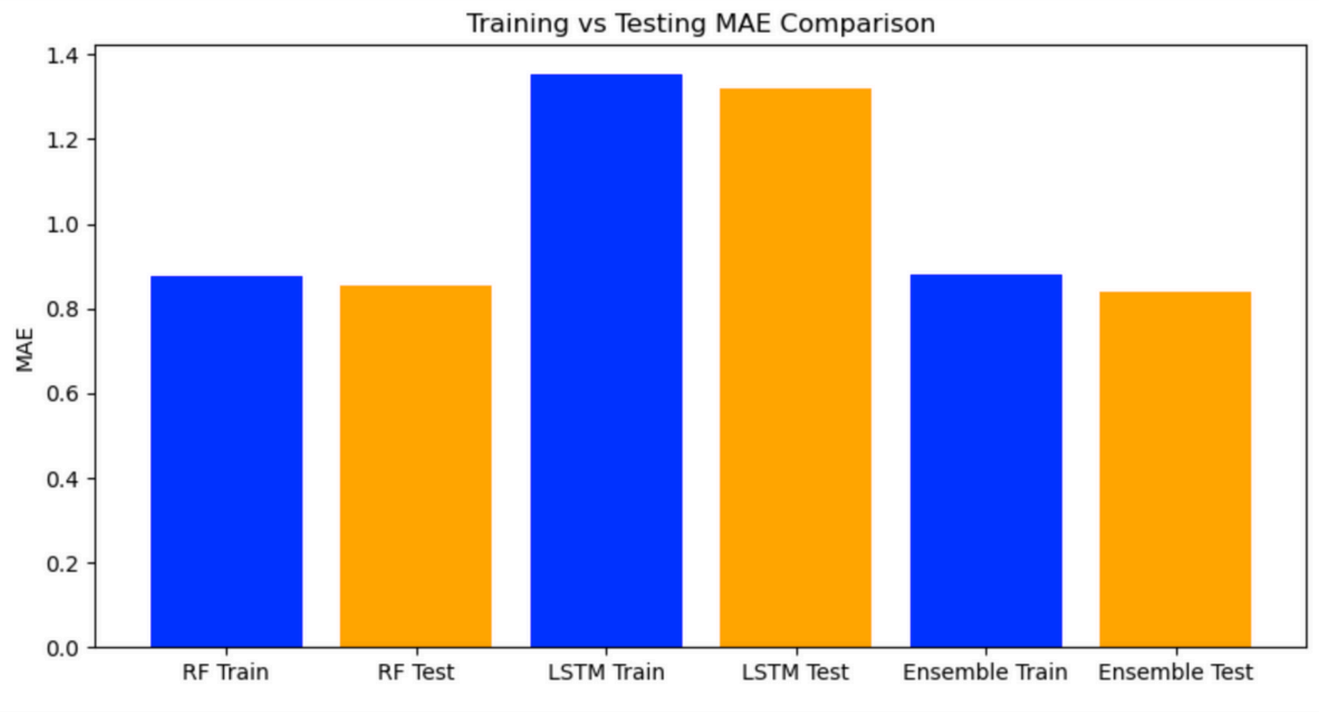
d) Ensemble Weighted Model:

- To address individual model weaknesses, an ensemble weighted by inverse MAE was created to leverage the strengths of Random Forest (strong generalization) and LSTM (temporal sequence learning).

- This approach ensured stability and minimized the errors by combining multiple models, which ultimately led to the best performance

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Models Used



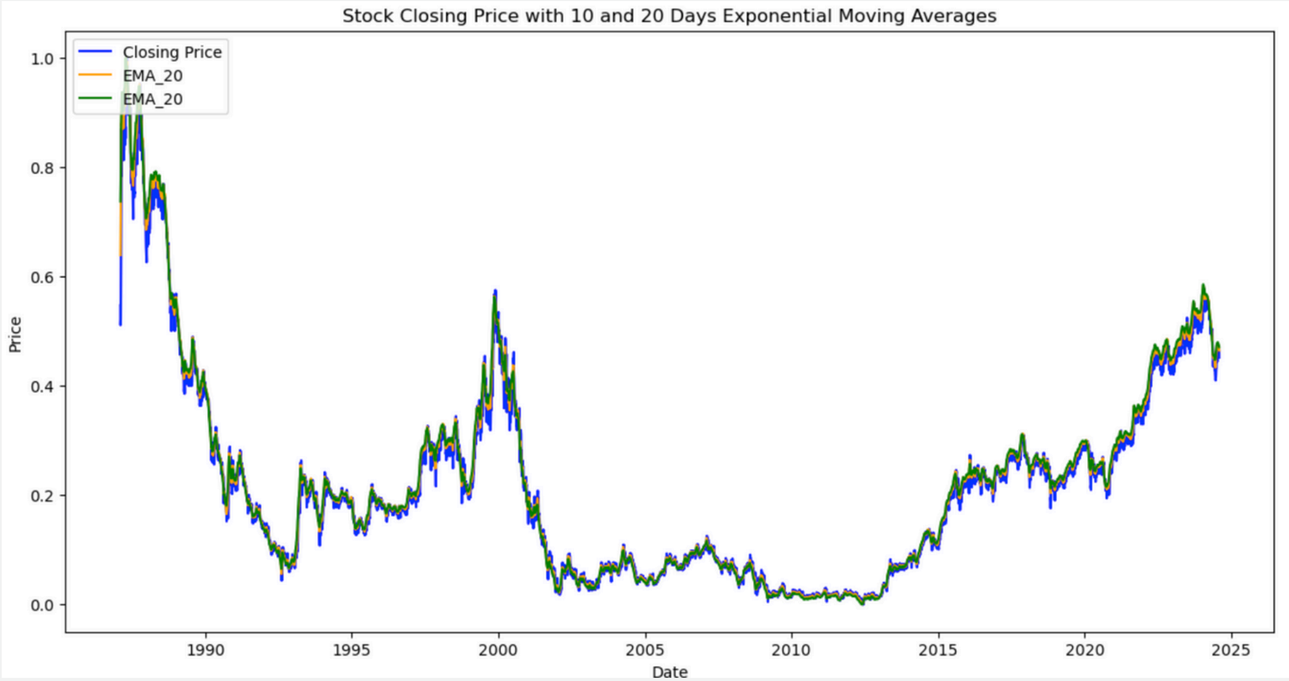
d) Ensemble Weighted Model:

- Ensemble Weighted Model MAE: 0.8412 (The lowest among all models).

  Random Forest Training MAE: 0.875703
  LSTM Training MAE: 1.354680
  Ensemble Weighted Training MAE: 0.880653
  Comparing Training vs Testing MAEs:
  Random Forest Testing MAE: 0.853645
  LSTM Testing MAE: 1.320600
  Ensemble Weighted Testing MAE: 0.841169

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Feature Engineering Methods and Rationale



a) Moving Averages (SMA_10,EMA_10,EMA_20)

- Used to capture long-term and short-term price trends while removing short-term noise, enabling the models to identify sustained price movements.

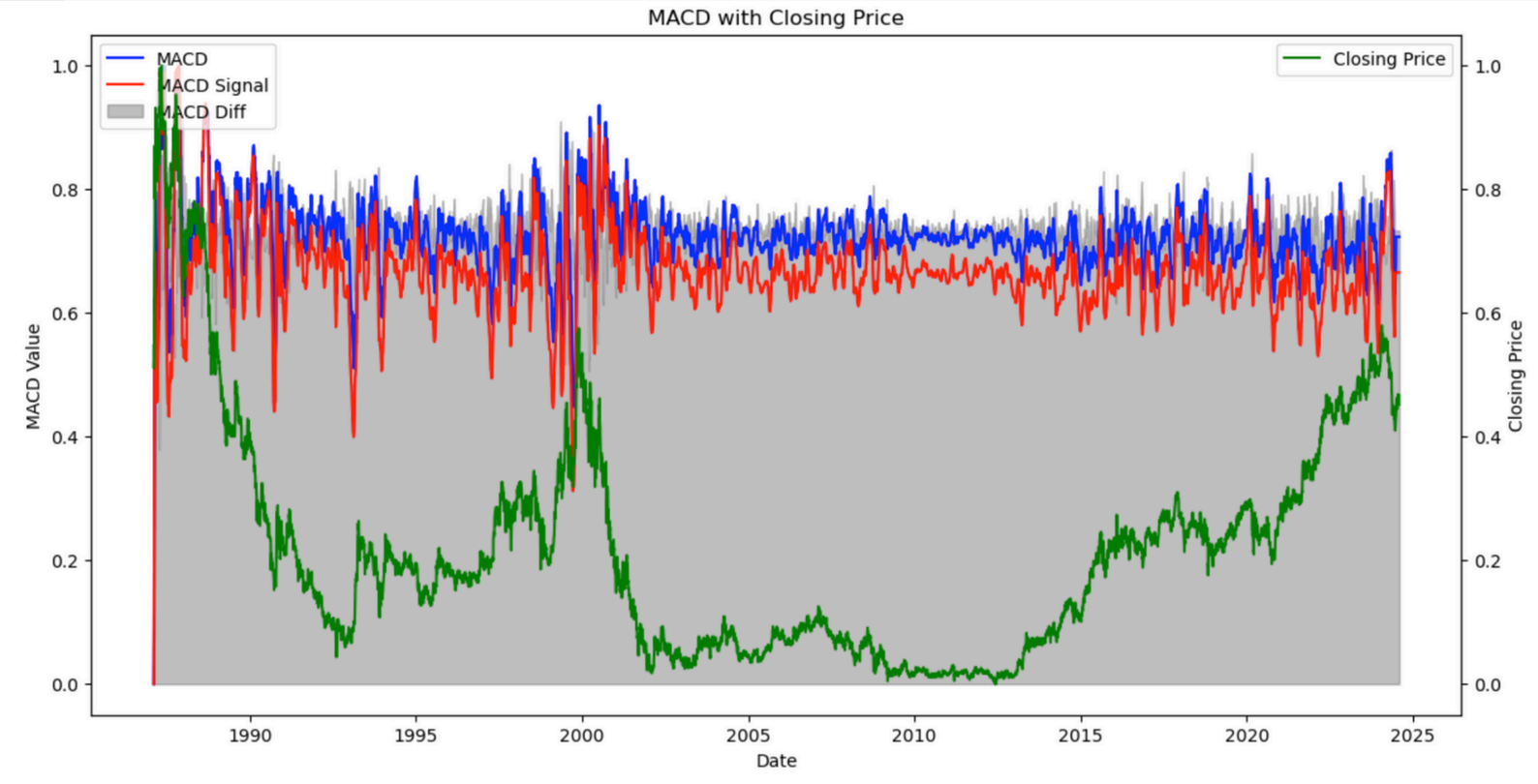
b)Technical Indicators

RSI (Relative Strength Index)

- To understand market momentum, which can indicate overbought or oversold conditions, allowing the model to better time predictions based on momentum.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Feature Engineering Methods and Rationale



b)MACD (Moving Average Convergence Divergence):

- Used to capture the relationship between two moving averages, helping to predict trend reversals and identify market trends.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Feature Engineering Methods and Rationale

a) Lag Features:

- Historical prices have a direct influence on the current stock value, so including lagged versions (e.g., Lag 1, Lag 2, Lag 3) of the closing price helps models learn relationships between past and present prices.
- This temporal correlation is crucial for models like Random Forest and LSTM to understand short-term dependencies.

a) Volume Analysis

- Volume can be an important indicator of market interest and potential price changes. Volume spikes often precede large price movements, making it a critical feature.
- Volume was log-transformed to deal with extreme skewness, normalizing it to improve model performance.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Feature Engineering Methods and Rationale

c) Date Features:

- Added features like Day of the Week to capture weekly cyclic patterns, and indicators for month start or holidays to address potential market behavior changes around such times.
- Such features were critical for models to learn cyclical patterns and adjust predictions based on market seasonality.

d) Correlation Analysis and Feature Selection

- To avoid multicollinearity and overfitting, features that were highly correlated with each other were carefully handled. Random Forest benefited from feature importance calculations, helping us select the most influential features.

KEY INSIGHTS FROM FEATURE ENGINEERING

IMPORTANCE OF LAG FEATURES	Lagged closing prices contributed significantly to both Random Forest and LSTM's ability to predict future prices.
IMPACT OF VOLUME	Including volume as a feature helped capture market activity spikes, which often precede price changes, contributing to model accuracy.
TECHNICAL INDICATORS ADDED PREDICTIVE POWER	The addition of RSI and MACD provided insights into market momentum, improving prediction in combination with traditional features.

EVALUATION INDICATORS AND MODEL PERFORMANCE ANALYSIS

Evaluation Indicators Used

- To thoroughly evaluate the predictive performance of each model, we used multiple evaluation metrics that provide unique perspectives on model effectiveness:

Model	MAE Value	Remark
Baseline Naïve	1.0834	This serves as the simplest benchmark to judge if advanced models are providing value
Random Forest	0.8536	Lower error, demonstrating its ability to effectively learn complex relationships
LSTM	1.3206	Reasonable performance but higher than Random Forest due to challenges in overfitting
Ensemble Weighted	0.8412	The best performance, reflecting effective integration of multiple models

a) Mean Absolute Error (MAE)

Represents the average magnitude of errors in predictions, without considering their direction. MAE is easy to interpret as it is expressed in the same units as the data.

Why it matter : MAE helped us determine which model consistently had the lowest deviation from the actual stock price. Lower MAE generally indicated better model performance in capturing the trends.

EVALUATION INDICATORS AND MODEL PERFORMANCE ANALYSIS

- The Ensemble Weighted Model achieved the **lowest MAE**, emphasizing the power of combining different strengths of Random Forest and LSTM models.

Model	MSE Value	Remark
Random Forest	1.2153	Robust learning of relationships between features, resulting in reduced deviations
LSTM	3.1013	Higher variance compared to Random Forest, suggesting it struggled with rapid price changes
Ensemble Weighted Model	1.2022	Slightly better performance than Random Forest, highlighting stability

b) Mean Squared Error (MSE)

Measures the average of the squares of errors. MSE penalizes larger errors more heavily, making it a good indicator for identifying models that have occasional large errors.

Why it matter : By squaring the errors, MSE highlights the impact of large deviations, thus helping in identifying models prone to occasional significant errors.

EVALUATION INDICATORS AND MODEL PERFORMANCE ANALYSIS

- The Ensemble Weighted Model achieved the **lowest MAE**, emphasizing the power of combining different strengths of Random Forest and LSTM models.

Model	RMSE Value	Remark
Random Forest	1.1024	Significantly lower, showing better adaptability
LSTM	1.7610	Good but not the best, implying challenges in learning across all sequences
Ensemble Weighted Model	1.0965	Slight improvement, indicating a combination approach works best

a) Root Mean Squared Error (RMSE))

The square root of MSE, providing the error metric in the same units as the target variable. RMSE penalizes larger errors and is more sensitive to outliers.

Why it matter : MSE allows us to understand the average size of the prediction error in a comparable way to the stock prices themselves, helping gauge overall model reliability.

- The **lower RMSE for Ensemble Model** compared to Random Forest shows how blending models can improve prediction reliability and reduce high-error occurrences.

EVALUATION INDICATORS AND MODEL PERFORMANCE ANALYSIS

a) Mean Absolute Percentage Error (MAPE)

Model	MAPE Value	Remark
Random Forest	0.7076%	Consistent and low, indicating strong relative accuracy
LSTM	1.0768%	Acceptable, but slightly higher compared to Random Forest
Ensemble Weighted Model	0.6887%	Best relative accuracy

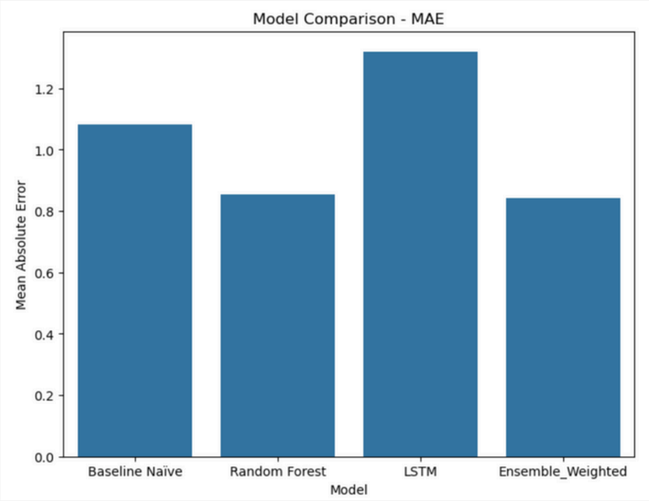
Measures prediction accuracy as a percentage, reflecting the size of the error in relation to the target values. Useful for relative comparison irrespective of scale.

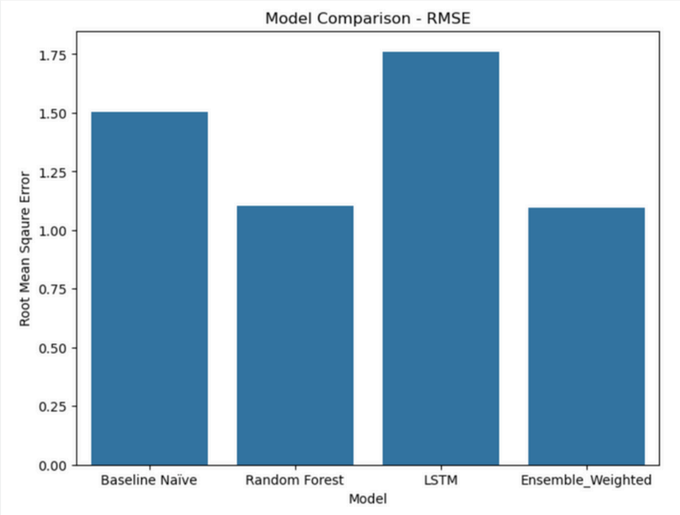
Why it matter : MAPE is an intuitive way to assess how well the model is performing in predicting different price ranges. It helps understand relative accuracy across fluctuating prices.

- The **low MAPE of the Ensemble Model** shows that it has the best accuracy, regardless of the varying scales of stock prices, proving its robustness across different price ranges.

MODEL SELECTION AND FEATURE ENGINEERING METHODS

Key Visual Insights





a) MAE and RMSE Comparisons:

Random Forest and Ensemble Weighted Models showed consistently lower MAE and RMSE values, emphasizing their ability to adapt to both short-term fluctuations and broader trends. The Ensemble Model achieved the lowest error, signifying its balanced learning from both Random Forest and LSTM.

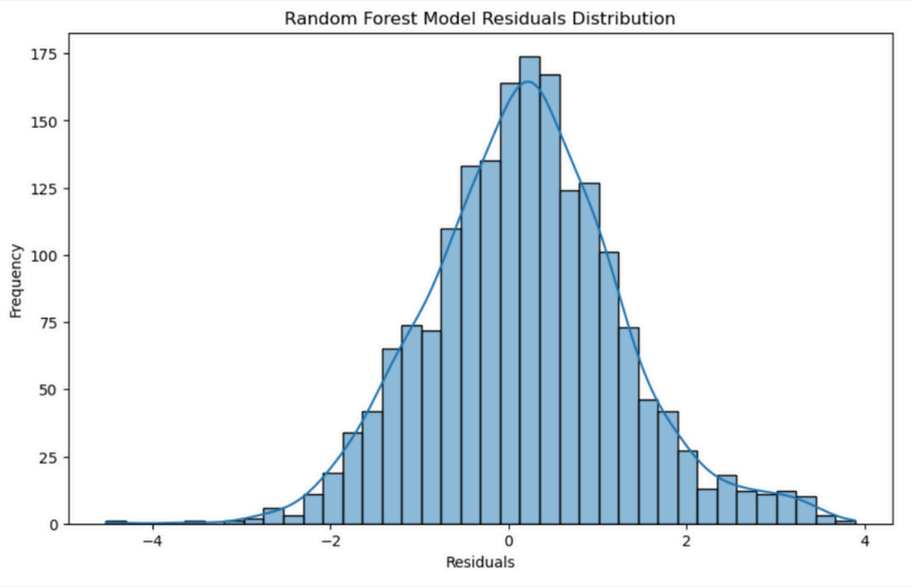
LSTM captured sequential dependencies but was more prone to overfitting, leading to slightly higher MAE and RMSE values compared to Random Forest.

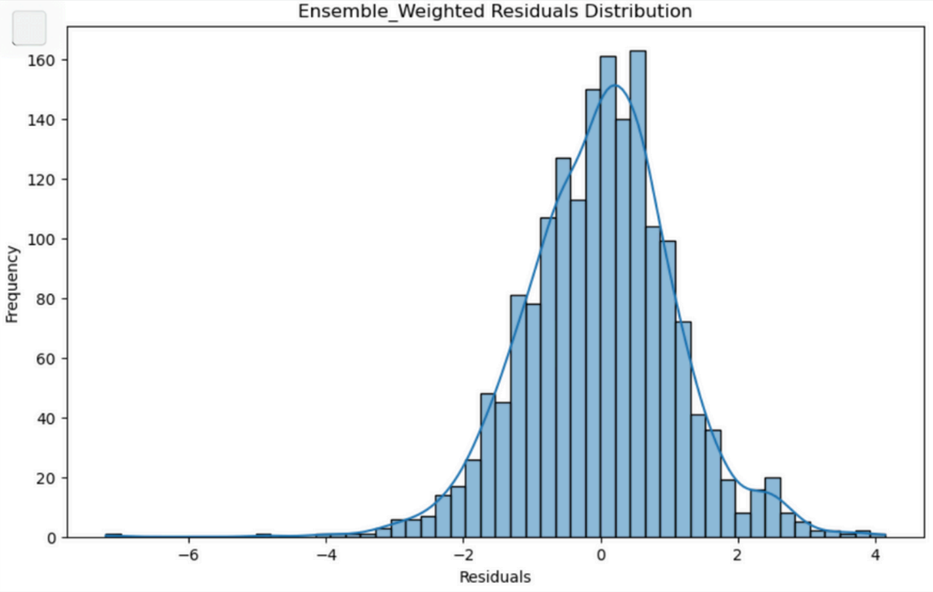
MODEL SELECTION AND FEATURE ENGINEERING METHODS

Key Visual Insights

b) Residual and Error Analysis

- Residuals for Random Forest and Ensemble Models had a more uniform distribution around zero, indicating minimal bias in predictions.





- Residual Distribution Plot for Random Forest and Ensemble - Shows tighter clustering around zero, confirming effective bias management.

OVERALL BEST MODEL

- The **Ensemble Weighted Model** consistently outperformed others across all metrics. Its combination approach reduced individual weaknesses (Random Forest's underestimation of certain trends, LSTM's susceptibility to overfitting) and boosted overall accuracy.
- This resulted in a **MAE of 0.8412** and a **MAPE of 0.6887%**, the lowest across all models, highlighting that combining diverse approaches can yield more stable and accurate predictions.

VERIFICATION CONTENT

Improvement measures were essential to enhance the accuracy and stability of our stock price predictions. We formulated hypotheses around model design, training techniques, and ensemble combinations.

Hypotheses Formulated

a) Hypothesis 1: Hyperparameter Tuning Improves Random Forest Performance

- **Hypothesis**: By carefully adjusting hyperparameters like `n_estimators`, `max_depth`, `min_samples_split`, and `max_features`, we hypothesized that Random Forest performance could be improved, resulting in lower prediction errors.
- **Verification Method:** We conducted extensive Grid Search with cross-validation to find the optimal values.
- **Result:** The best hyperparameters (`max_depth: 5`, `max_features: 0.3`, etc.) were obtained, and the Random Forest's MAE was reduced to **0.8536**, a noticeable improvement over default settings

VERIFICATION CONTENT

b) Hypothesis 2: Incorporating Dropout and Regularization in LSTM Improves Generalization

- **Hypothesis:** Adding Dropout layers and L2 Regularization to the LSTM model would reduce overfitting and improve generalization, lowering the MAE for the testing dataset.
- **Result:** The LSTM model with these improvements achieved an MAE of **1.3206**, which was lower compared to a non-regularized model.

c) Hypothesis 3: Weighted Ensemble of LSTM and Random Forest Will Provide the Best

- **Hypothesis:** An ensemble model, using inverse MAE weighting for LSTM and Random Forest predictions, would leverage the strengths of both and minimize weaknesses.
- **Result:** The LSTM model with these improvements achieved an MAE of **1.3206**, which was lower compared to a non-regularized model.

VERIFICATION RESULT 1

Hypothesis: We hypothesized that tuning hyperparameters like `n_estimators`, `max_depth`, `min_samples_split`, etc., in the Random Forest model would lead to improved prediction accuracy**.**

Verification Approach: An ensemble model, using inverse MAE weighting for LSTM and Random Forest predictions, would leverage the strengths of both and minimize weaknesses.

- Grid Search Cross-Validation :Conducted on the Random Forest model, varying parameters systematically.
- **Key Metrics Monitored**: Mean Absolute Error (MAE) and Mean Squared Error (MSE) during training and testing phases.

- **Best Parameters Identified:**
  - Max Depth: 5
  - Max Features: 0.3
  - Min Samples Split: 10
  - Number of Estimators: 50

Verification Results

- **MAE Reduction:** The MAE for the Random Forest dropped to 0.8536, compared to higher errors without tuning.

VERIFICATION RESULT 2

Testing Generalization: The Testing MAE (0.8536) was similar to the Training MAE (0.8956), indicating effective generalization without overfitting

Considerations:

Impact of Hyperparameter Tuning: The decrease in both training and testing error highlights that optimal parameter settings can significantly improve a model's learning capacity without leading to overfitting

Performance Limitation: Further reducing `max_depth` might have improved generalization, but could also risk underfitting. A balance was achieved with a depth of 5.

Verification Results and Considerations for Hypothesis 2

Hypothesis 2 Recap : We hypothesized that incorporating dropout layers and L2 regularization in the LSTM architecture would help **prevent overfitting** and **improve generalization** to unseen data.

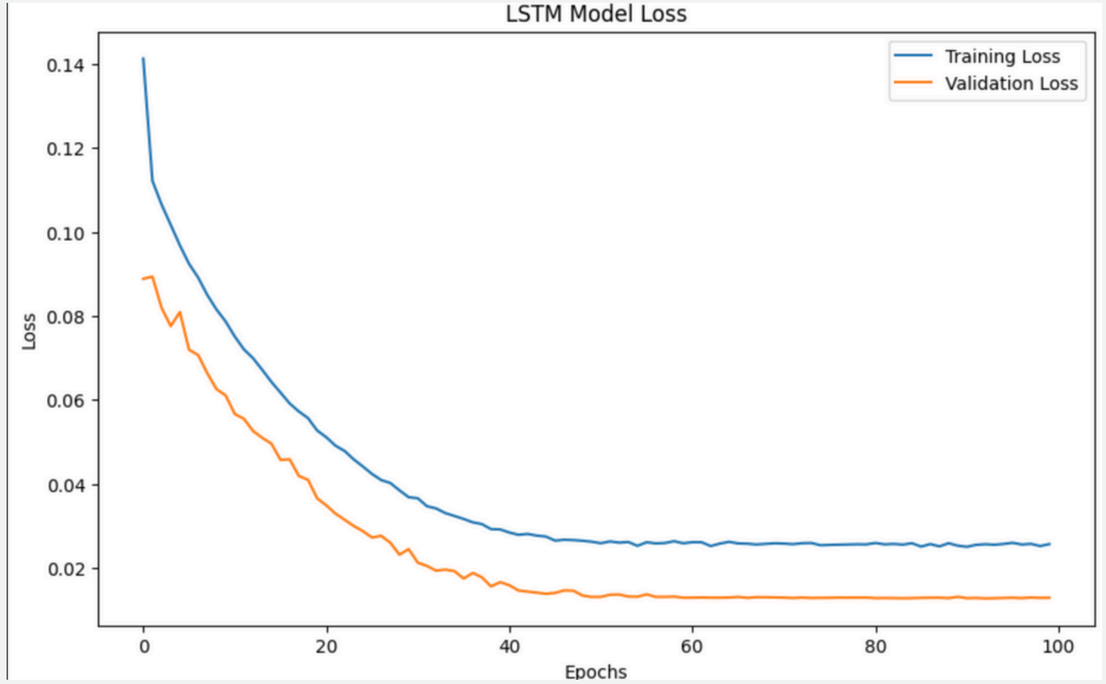
Verification Approach :

- Used **L2 Regularization** in the dense output layer to add a penalty for high weights.
- Applied **Early Stopping** to terminate training when the validation loss stopped improving.
- **Training and Validation Observations:**
- Early stopping was triggered after around **20 epochs**, with both training and validation loss converging.

VERIFICATION RESULT 2

Verification Results

- **Testing MAE**: The LSTM model achieved a final testing MAE of 1.3206.
- Training vs. Validation Loss: Both curves followed a similar downward trend, suggesting improved generalization.
- **Regularization Effectiveness**: L2 Regularization helped prevent the model from memorizing the training data, while dropout induced additional randomness, encouraging better generalization.



Considerations:

- **Overfitting Mitigation**: Incorporating dropout and L2 regularization **effectively balanced learning**, especially in a sequence model like LSTM, which tends to memorize sequential data.

- **Potential Improvements**: Testing with a **different dropout rate** (e.g., 0.3 or 0.4) and a larger LSTM layer size could provide additional improvement opportunities.

VERIFICATION RESULT 3

Hypothesis : We hypothesized that combining the LSTM and Random Forest models in a **weighted ensemble**, using inverse MAE as weights, would **yield better predictive accuracy** compared to individual models by leveraging the complementary strengths of each.

Verification Approach

Ensemble Construction :

- The **Random Forest** model excelled at capturing non-linear relationships, while the **LSTM** model effectively captured sequential dependencies.
- We calculated the weights for each model based on their i**nverse MAE**. This gave more influence to the model with lower MAE (Random Forest).

Weights Used:

- R**andom Forest**: 0.6074
- **LSTM**: 0.3926

Weighted Averaging:

- The final predictions were computed by taking a weighted average of the individual model predictions, leading to an ensemble output.

.

VERIFICATION RESULT 3

Verification Results

- **Ensemble MAE**: **0.8412**, which was lower than the individual **Random Forest (0.8536)** and **LSTM (1.3206)** models.

  MAPE and RMSE:

  - **Mean Absolute Percentage Error (MAPE):** 0.6887%, indicating the smallest percentage error across different stoc price ranges.
  - **- Root Mean Squared Error (RMSE):** 1.0965, reflecting improved consistency in predictions.

  Generalization Performance

  - Training and testing errors showed a similar pattern, indicating robust performance without significant overfitting.

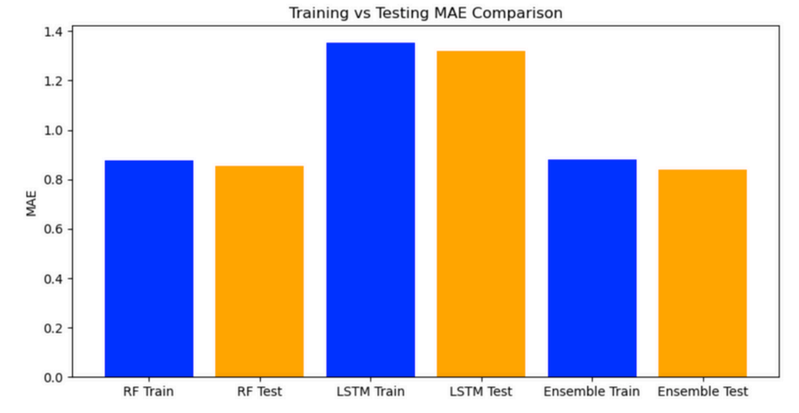
  - **Training MAE**: 0.8807
  - **Testing MAE**: 0.8412

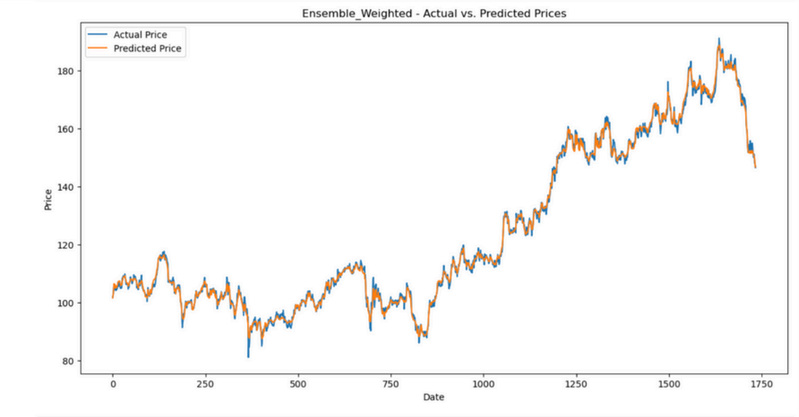
Considerations

- **Ensemble Effectiveness**: The Ensemble Weighted Model performed better because it balanced the sequential learning of LSTM with the non-linear adaptability of Random Forest, mitigating each model's individual weaknesses.

VERIFICATION RESULT 2

- **Reduced Overfitting :** Compared to LSTM alone, the ensemble model effectively reduced overfitting, evident in the lower error metrics and similar training vs. testing MAE.





- **Practical Application:** For real-world scenarios, such a combined model offers better generalization, meaning it’s less likely to fail during extreme market conditions, thus providing more dependable predictions.

Insights from Weighted Ensemble

- **Diversity in Learning:** The success of the ensemble lies in combining diverse perspectives—Random Forest for non-linear relationships and LSTM for sequential patterns**.**

- **Practical Application:** For real-world scenarios, such a combined model offers better generalization, meaning it’s less likely to fail during extreme market conditions, thus providing more dependable predictions.

SUMMARY OF RESULTS

1.Model Performance Overview

- We evaluated five different models for predicting NTT stock prices, including Baseline Naïve, ARIMA, Random Forest, LSTM, and an Ensemble Weighted Model.
- **Ensemble Weighted Model e**merged as the best performer, achieving the **lowest Mean Absolute Error (MAE) of 0.8412,** the lowest Mean Absolute Percentage Error (MAPE) of 0.6887%, and the lowest RMSE of 1.0965.

2. Key Insights:

Random Forest vs. LSTM

- Random Forest performed better in capturing non-linear dependencies, resulting in a lower MAE (0.8536) compared to LSTM (1.3206).
- LSTM, despite its sequential learning capacity, showed a tendency to overfit, although applying dropout and regularization reduced this issue to some extent.

Ensemble Effectiveness

- By combining Random Forest and LSTM, the Ensemble Model successfully reduced individual weaknesses, resulting in improved prediction accuracy and stability.
- The weighted approach allowed us to capitalize on both sequential dependencies (from LSTM) and non-linear adaptability (from Random Forest).

SUMMARY OF RESULTS

Residual and Error Distribution

- Ensemble Model Residuals were more uniformly distributed around zero, reflecting unbiased predictions, and exhibited the smallest error variance compared to other models.

3. Evaluation Indicators:

- Metrics used included **MAE, MSE, RMSE, and MAPE.**
- The **Ensemble Weighted Model** consistently outperformed others in each metric, suggesting robustness in varying market conditions.

FUTURE OUTLOOK

1.Model Enhancements

Incorporate Transformer-Based Models

- Given the success of Transformer architectures in sequential tasks, experimenting with Transformer models for time series prediction could enhance learning of complex temporal dependencies. Transformers offer advantages like better long-range pattern recognition, which could potentially outperform LSTMs.
- The hypothesis is that a Transformer-based model may adapt better to sudden market shifts, given its self-attention mechanism.

2. Data Expansion

Broaden the Dataset

- Including additional features such as macroeconomic indicators, company financials, or market sentiment data (e.g., news or social media mentions) can help capture broader influences on stock price movement.
- Seasonal and Volatility Analysis: Introducing additional seasonal indicators or volatility indexes might improve the model's predictive accuracy, especially during high fluctuation periods.

FUTURE OUTLOOK

3. **Ensemble Learning Expansion**

   Diversify the Ensemble :

   - Stacking Ensemble**: Experimenting with **stacking** (a multi-layered ensemble approach) may further improve accuracy by allowing higher-level models to learn from the outputs of base models.

4. **Practical Applications**

   Deploying as a Real-Time Prediction Tool :

   - The final ensemble model can be deployed for real-time stock price predictions, where daily retraining ensures that the model stays updated with the latest data trends.

   User Interface for Insights :

   - Building a user-friendly dashboard that provides real-time predictions and explanations (using SHAP values for feature importance) could be valuable for investors and analysts to understand model predictions better.

THANK YOU

Presentation by Saurav Raj