Model Selection & Justification Report

1. Models Tested

To predict the stock closing price **5 days into the future**, we tested multiple models with different characteristics:

1.1 Linear Regression

- **Type:** Simple regression-based model
- **Why?**
  - Provides a **baseline model** for stock price prediction.
  - Assumes a **linear relationship** between features and the target price.
  - Easy to interpret and computationally efficient.

1.2 Random Forest Regressor

- **Type:** Ensemble-based decision tree model
- **Why?**
  - Captures **non-linear relationships** in stock price movements.
  - Handles missing data and noisy features better than linear models.
  - Often performs well on structured time-series data.

---

2. Evaluation Metrics Used

To compare the models, we used **Root Mean Square Error (RMSE)** as the primary evaluation metric.

2.1 Why RMSE?

- RMSE penalizes large errors more than small ones, making it **sensitive to extreme values**.
- **Formula:**

$$RMSE = \sqrt{\frac{1}{n}\sum(y_{\text{actual}} - y_{\text{predicted}})^2}$$

- **Interpretation:**
    - **Lower RMSE → Better model performance** (lower prediction error).
    - **Higher RMSE →** Model struggles to fit the data properly.

2.2 Other Potential Metrics (Not Used Yet)

- **Mean Absolute Error (MAE)** → Measures average error magnitude, less sensitive to outliers.
- **R^2 Score (Coefficient of Determination)** → Shows how well the model explains variance in the stock price.

---

3. Results: Model Performance Comparison

Model	RMSE (Lower is Better)
Linear Regression	**5.06**
Random Forest	28.63

Key Observations

- **Linear Regression performed significantly better**, achieving an RMSE of **5.06**.
- **Random Forest performed poorly (RMSE = 28.63)**, possibly due to overfitting or high variance in stock data.

---

4. Final Model Selection & Justification

After evaluating both models, we selected **Linear Regression** as the final model.

4.1 Why Linear Regression?

Better RMSE: It achieved the lowest error (5.06).
Simplicity & Interpretability: Linear models are easy to understand.
Stable Performance: Less prone to overfitting compared to complex tree-based models.

4.2 Why Not Random Forest?

High RMSE (28.63): Indicates poor generalization on stock price movements.
Overfitting Risk: Tree-based models may struggle with financial time series.
Less Explainable: Harder to interpret compared to linear models.

---

5. Model Limitations & Future Improvements

5.1 Current Limitations

No External Market Factors Considered

- Our model only uses historical stock prices and technical indicators.
- Real-world stock prices are influenced by **news, economic reports, interest rates, and global events**.

Linear Model Assumptions May Not Hold

- Stock prices **don't always follow linear relationships**.
- Market movements can be **highly non-linear**, especially during crashes or booms.

Short-Term Prediction Focus

- The model predicts **only 5 days ahead**.
- Long-term stock price trends might require **LSTM-based deep learning models**.

---

6. Future Improvements

6.1 Advanced Time-Series Models

LSTM (Long Short-Term Memory Networks)

- Captures long-term dependencies in stock price movements.
- Better suited for **sequence-based predictions**.
- Requires **more historical data** for optimal performance.

XGBoost or LightGBM

- Tree-based models with better regularization than Random Forest.
- Can handle **non-linear relationships** while maintaining efficiency.

6.2 Incorporating External Data

News Sentiment Analysis

- Use **Natural Language Processing (NLP)** to analyze news headlines.
- Example: **Positive news → Price rise, Negative news → Price drop.**

Macroeconomic Indicators

- Consider variables like **inflation, interest rates, and GDP growth**.
- These factors influence stock prices beyond historical patterns.

Market Trends & Technical Indicators

- Add **Bollinger Bands, MACD, and Fibonacci Retracements** for more sophisticated trading insights.