Model Selection Documentation

1. Model Training and Evaluation

Following the comprehensive exploratory data analysis, multiple modeling approaches were tested to identify the most effective predictive model for the financial time series data. This document outlines the model selection process, evaluation metrics, and justification for the final model choice.

1.1 Models Tested

A diverse range of models were implemented and evaluated:

Model Type	Description
Linear Regression	Simple but effective linear approach
Random Forest	Ensemble method using multiple decision trees
XGBoost	Gradient boosting implementation with regularization
Decision Tree	Single tree-based approach
KNN	K-Nearest Neighbors regression
SVM	Support Vector Machine regression
ARIMA	AutoRegressive Integrated Moving Average
SARIMA	Seasonal ARIMA for time series with seasonality
LSTM	Long Short-Term Memory neural network
RNN	Recurrent Neural Network for sequence data

1.2 Evaluation Methodology

Each model was trained on the chronologically ordered training set and evaluated on the test set. To ensure reliable assessment, predictions were visualized alongside actual test values to identify patterns in forecast errors and model performance across different market conditions.

2. Evaluation Metrics

Two primary metrics were used to quantitatively evaluate model performance:

2.1 R² Score (Coefficient of Determination)

The R² score measures the proportion of variance in the dependent variable that can be explained by the independent variables. It provides insights into how well the model captures the overall trend and pattern in the data.

- $R^2 = 1$: Perfect prediction
- $R^2 = 0$: Model performs no better than using the mean value
- $R^2 < 0$: Model performs worse than using the mean value

2.2 Root Mean Squared Error (RMSE)

RMSE measures the average magnitude of prediction errors, giving greater weight to larger errors due to the squaring operation:

- Lower RMSE values indicate better model performance
- RMSE is sensitive to outliers and large errors
- RMSE is in the same units as the target variable, making it interpretable

3. Model Comparison Results

3.1 Performance Summary

After thorough evaluation, the following performance trends were observed:

1. **Linear Regression** demonstrated the highest overall performance with the best combination of R² score and lowest RMSE. Predictions from this model aligned closely with actual values across various market conditions.`(RMSE: 0.6029463971680233,R-squared: 0.9994674959091102)`
2. **ARIMA**, **SARIMA**, and **RNN** models showed promising results but performed slightly less effectively than Linear Regression. These models captured temporal dependencies well but occasionally struggled with extreme market movements.
3. **LSTM** showed potential for capturing complex patterns but required more data and tuning to reach optimal performance.
4. **Tree-based models** (Random Forest, Decision Tree, XGBoost) performed moderately well but showed signs of overfitting to the training data.
5. **KNN** and **SVM** demonstrated lower performance compared to other approaches.

3.2 Visual Comparison

Visualization of predictions against actual values confirmed the quantitative metrics. Linear Regression produced predictions that tracked actual market movements most accurately, with particularly strong performance during periods of moderate volatility.

4. Final Model Selection: Linear Regression

4.1 Justification

Linear Regression was selected as the final model based on several factors:

1. **Superior Performance**: Highest R^2 score and lowest RMSE among all tested models
2. **Consistency**: Reliable predictions across different market conditions
3. **Interpretability**: Clear relationship between features and predictions
4. **Computational Efficiency**: Faster training and prediction compared to more complex models
5. **Generalization**: Good performance on out-of-sample data, suggesting robust generalization

4.2 Implementation Details

The final Linear Regression model uses the following features:

- Historical price data (Open, High, Low, Close)
- Volume information
- Technical indicators derived from the raw data

5. Model Limitations and Potential Improvements

5.1 Identified Limitations

Despite its strong performance, the Linear Regression model has several limitations:

1. **Assumption of Linearity**: May not capture complex non-linear relationships in the data
2. **Limited Memory**: Does not explicitly model long-term dependencies in time series
3. **Sensitivity to Outliers**: Extreme market events can disproportionately influence the model
4. **Feature Dependency**: Relies heavily on the quality of engineered features

5.2 Potential Improvements

With additional time and resources, the following improvements could be implemented:

1. **Hybrid Modeling Approach**: Combine Linear Regression with ARIMA/SARIMA components for both trend and seasonal patterns
2. **Feature Engineering**: Develop more sophisticated technical indicators and external features
3. **Ensemble Methods**: Create an ensemble of the top-performing models to improve robustness
4. **Deep Learning Optimization**: Further tune LSTM/RNN architectures with larger datasets
5. **Market Regime Detection**: Implement regime-switching models to adapt to changing market conditions
6. **Alternative Data Integration**: Incorporate sentiment analysis, macroeconomic indicators, and other alternative data sources

5.3 Final Approach

The final approach combines the Linear Regression model for primary predictions with ARIMA model outputs as a complementary source. Results from both models were exported to a CSV file containing:

- Actual values

- Linear Regression predictions
- ARIMA predictions
- Prediction errors and performance metrics

This dual-model approach provides a more comprehensive view of potential future price movements and serves as a foundation for further model development and refinement.

Data_Dominators Task 4 part2