An Approach to Stock Prediction on Japanese Stock Data

Prepared by - Satyam Agrawal

Student of Artificial Intelligence

National Institute of Technology Karnataka

Surathkal, Karnataka, India 575025

11-10-2024

Table Of Contents

Background

- **Market Dynamics:** Stock markets are highly volatile, and predicting stock movements can offer significant financial advantages.
- **Economic Impact:** Accurate stock predictions support investors, firms, and governments in making informed decisions about investment strategies, risk management, and economic policies.
- **AI in Finance:** With the rise of artificial intelligence and machine learning, stock prediction is one of the key areas where data-driven insights can provide a competitive edge.

Data Preprocessing

- Converted Volume and Change % to numeric format for accurate analysis.
- Reversed the data order to maintain a time series flow from past to present.

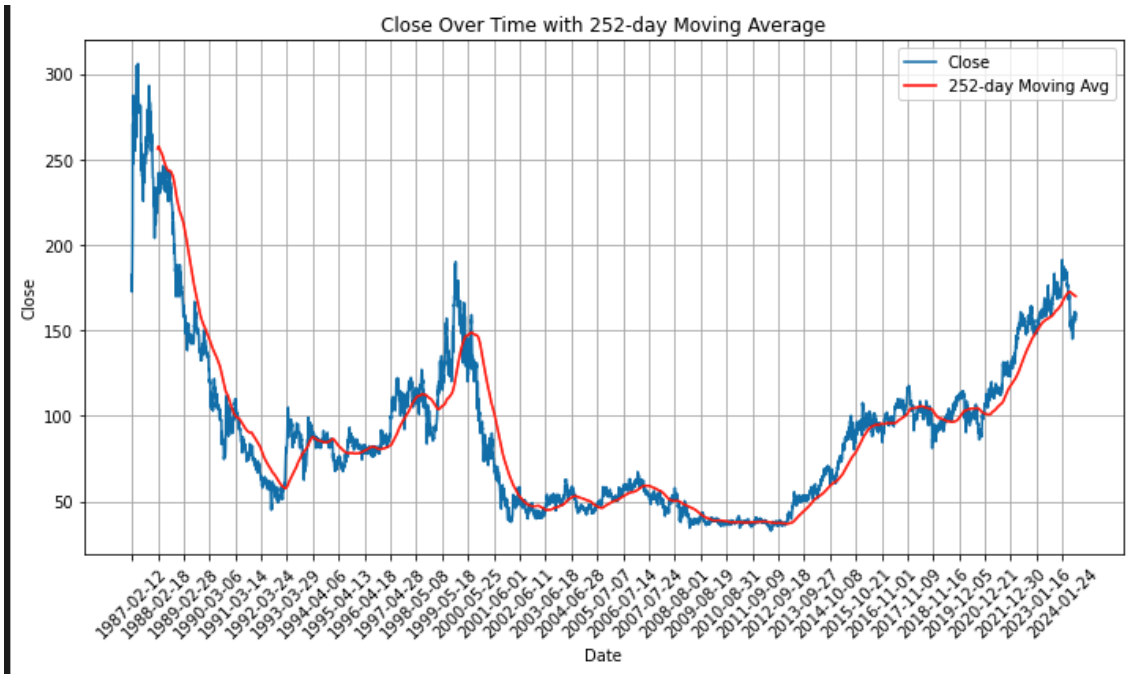
Summary Statistics

- Descriptive statistics show typical values for Open, High, Low, Close, and Volume.
- No missing values were found.

	Close	Open	High	Low	Volume	Change %
count	9202.000000	9202.000000	9202.000000	9202.000000	9202.000000	9202.000000
mean	92.180961	92.256183	93.176451	91.330146	172.667670	0.017502
std	50.452228	50.598215	51.049837	50.087405	125.127965	1.876667
min	33.000000	33.000000	33.200000	32.200000	9.340000	-14.740000
25%	52.000000	52.100000	52.800000	51.500000	80.730000	-0.940000
50%	85.100000	85.100000	86.050000	84.200000	154.015000	0.000000
75%	110.800000	110.800000	111.900000	109.275000	230.522500	0.900000
max	305.900000	309.800000	311.800000	303.900000	1280.000000	16.250000

Time Series Plots with Moving Average

- Open, High, Low, and Close prices show similar trends and high correlation.
- Market instability was evident between 1987 and 2001.
- Volume traded peaked in 2013-14 and declined steadily afterward.



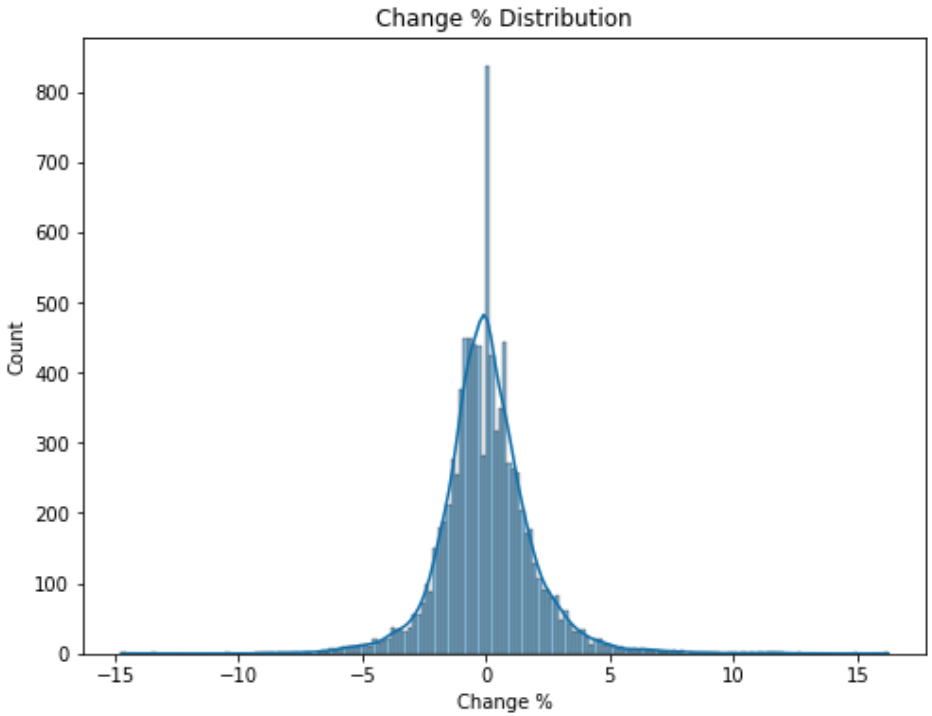
Correlation Matrix

- Strong correlation among Open, High, Low, and Close prices.
- Volume negatively correlated with prices, indicating more trades at lower prices.
- Change % has no significant correlation with prices but shows a slight positive correlation with Volume.

	Close	Open	High	Low	Volume	Change %
Close	1.000000	0.999547	0.999795	0.999754	-0.317508	0.017166
Open	0.999547	1.000000	0.999755	0.999791	-0.320279	-0.002860
High	0.999795	0.999755	1.000000	0.999660	-0.315402	0.007792
Low	0.999754	0.999791	0.999660	1.000000	-0.322392	0.005815
Volume	-0.317508	-0.320279	-0.315402	-0.322392	1.000000	0.130238
Change %	0.017166	-0.002860	0.007792	0.005815	0.130238	1.000000

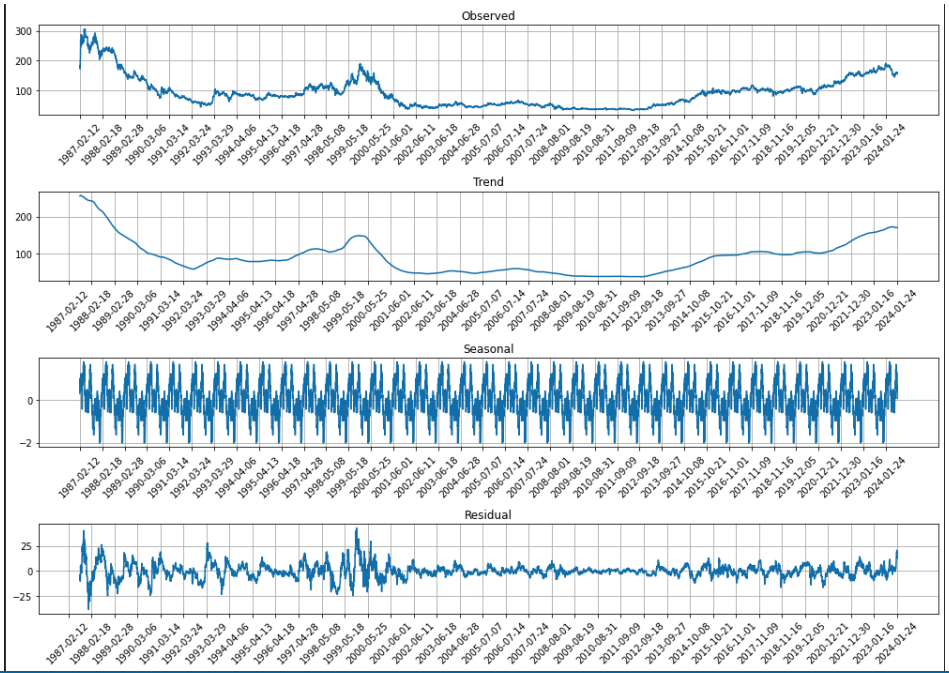
Distribution Analysis

- Most data points followed a normal distribution, especially for Change %, which fits common stock market assumptions.



Seasonal Decomposition

- Trend: Market experienced periods of price fall (1987-92), stabilization (93-99), and another fall (99-01).
- Seasonality: Not significant, showing small deviations.
- Residual: Highlights periods of instability with large swings during 1987-2001, followed by stability.



Anomaly Detection

- Used a 504-day window and a Z-score threshold of 4 for anomaly detection.
- Key anomalies found:
  - May 1997: Pre-Asian Financial Crisis.
  - 2008: Global Financial Crisis.
  - 2013: Major rise in the Japanese market (57% growth).

Feature Creation and Extraction

- Various lag features, moving averages, volatility measures, and momentum indicators are created.
- Additional features such as Relative Strength Index (RSI), Volume Weighted Average Price (VWAP), and Sharpe Ratio are computed.
- The target variable is set to the next day's closing price.

Data Splitting and Scaling

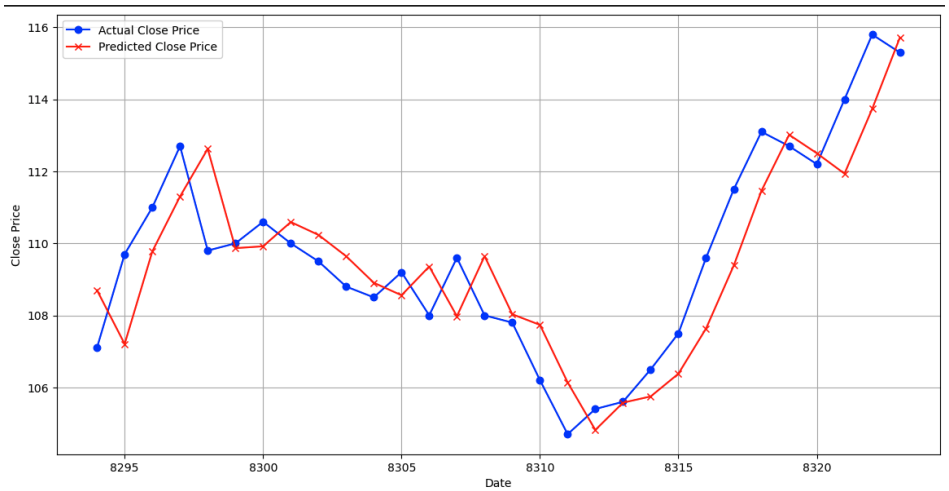
- The dataset is split into training and testing sets based on a specified date.
- The features are scaled using both `StandardScaler` and `MinMaxScaler`.

Modeling

- Multiple regression models are instantiated, trained, and evaluated on the test data.
- The aim of the exercise is to identify the best performing simple ML model to use it to derive the best set of Features (Feature Selection)
- Performance metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are computed.
- A comparison plot of the models is generated to visualize performance.

Problem Faced

- Metrics were unable to determine model performance due to which feature selection could not be performed.
- While Linear Regression based models had the best RMSE and MAE, upon closer inspection, it was found that they were simply presenting last day's value for Close Price as the predicted price.



Classification Approach

- Due to the previously mentioned issue, the problem is reframed as a classification task, predicting whether the price will go up or down.
- Mutual information is used for feature selection in the classification model.
- All features who have MI greater than 0 are selected and we move to the model selection and training step.

Model comparison using Classification Task

- Multiple models were compared for the task of classification, so as to predict if the price goes up or down.
- A separate comparison with Post-2000 data was also carried out to test the Hypothesis presented initially.

Challenges and Proposed Solutions

- **Poor Model Performance:** AUC, F1 Score, and Accuracy were all approximately 50%, indicating ineffective classification.
- **Bias in Recall Values:** The recall for the positive class was near zero, suggesting potential bias stemming from a declining market.
- **Post-2000 Data Issues:** Challenges arose due to discrepancies in data scale, where test data exhibited a higher range than training data.
- Addressed issues by focusing solely on ”change-based” variables, leading to more stable predictions.
- Proposed a shift to regression modeling for predicting returns, essentially focusing on percentage changes in prices.

Return-Based Regression Approach

- Multiple models were tested for performance.
- The same methodology applied to the post-2000 dataset.
- Observed poor results for MAE, MAPE, and RMSE across models.
- Difficulty in selecting the best-performing model.
- LSTM was ultimately chosen for further development.
- LSTM underwent additional fine-tuning to enhance performance.

Reason for Choosing LSTM

- **Lack of Clear Leader:** No single model demonstrated consistent superior performance.
- **Flexibility for Modification:** LSTM offers extensive opportunities for customization and tuning.
- **Proven Efficacy:** LSTM has been widely recommended for time series prediction tasks, making it a strong candidate for this analysis.

Possible Modifications for LSTM Fine-Tuning

- **Loss Function:** Change from Mean Squared Error (MSE) to Mean Absolute Error (MAE) for more robust performance against outliers.
- **Complex Architecture:** Increase the number of layers to capture more intricate patterns in the data.
- **Bidirectional LSTM (Bi-LSTM):** Experiment with a Bi-LSTM architecture to capture dependencies in both directions.
- **Activation Functions:** Explore different activation functions to enhance model learning capacity.
- **Increase Time Steps:** Adjust the number of time steps to provide the model with more historical context for predictions.

Chosen modifications resulted in an improvement of MAPE by 0.07%.

Evaluation Metrics - Classification

- **Accuracy:** Measures the proportion of correctly predicted instances out of the total instances. Used to assess overall model performance.
- **F1 Score:** The harmonic mean of precision and recall, balancing false positives and false negatives. Useful when classes are imbalanced.
- **Area Under the Curve (AUC):** Represents the ability of the model to distinguish between classes. A higher AUC indicates better model performance.

Evaluation Metrics - Regression

- **Root Mean Square Error (RMSE):** Measures the average magnitude of the errors between predicted and actual values, giving more weight to larger errors. Used for evaluating model accuracy.
- **Mean Absolute Error (MAE):** Calculates the average absolute errors between predicted and actual values. It provides a straightforward measure of prediction accuracy.
- **Mean Absolute Percentage Error (MAPE):** Expresses the accuracy as a percentage, allowing for easier interpretation across different scales. It helps assess model performance in percentage terms.

Conclusion

- Predicting stock market movements is inherently challenging, and the results from this project indicate that current models did not achieve satisfactory performance, highlighting the need for more advanced methodologies.
- The normal distribution of percentage changes reflects the inherent randomness in the market, complicating prediction tasks.
- Additionally, the lack of seasonality further contributes to the difficulties faced in modeling.
- Given these challenges, a refinement of the approach is necessary, incorporating more nuanced, domain-informed strategies to enhance predictive accuracy.