DSC 478 || Programming Machine Learning Applications

Final Project

Stock Market Forecasting using Machine Learning & Sentiment Analysis

Introduction

This project explores the predictive power of historical stock data and social media sentiments on stock prices, with a focus on Apple Inc. The objective is to leverage machine learning models to forecast stock price movements and evaluate the impact of social media sentiments on prediction accuracy.

Literature Review

Theoretical Background of Stock Market Prediction

Stock market prediction has long been a topic of interest among investors, researchers, and financial analysts. Traditional theories, such as the Efficient Market Hypothesis (EMH), suggest that stock prices reflect all available information and are thus unpredictable. However, behavioural finance introduces the concept that psychological factors and investor sentiment can influence market movements, providing a basis for the predictive modelling of stock prices.

Machine Learning Models in Financial Markets

The advent of machine learning has offered new methodologies for predicting stock market trends. Studies by Kumar and Thenmozhi (2006) and Patel et al. (2015) demonstrate the application of various machine learning models, such as Linear Regression, Random Forest, and neural networks, to predict stock prices with varying degrees of success. These models leverage historical stock data, including price and volume, to forecast future price movements.

- **Linear Regression** has been widely used for its simplicity and interpretability in modelling the linear relationship between stock market indicators and future stock prices (Tsai and Wang, 2009).
- **Random Forest Regression**, a more complex model, is praised for its ability to handle nonlinear relationships and feature interactions, making it particularly effective in financial applications where market conditions are dynamic and multifactorial (Krauss et al., 2017).
- **Gated Recurrent Units (GRU),** a type of recurrent neural network (RNN) architecture introduced by Cho et al. (2014). GRU has been designed to solve the vanishing gradient problem associated with traditional RNNs, making it highly effective for sequential data processing. The key feature of GRU is its gating mechanism, which controls the flow of information to be remembered or forgotten, thereby enhancing the model's ability to learn from long-term dependencies in time series data.

Application of GRU in Stock Market Prediction

The application of GRU models in financial markets is a growing area of research. These models can process sequential stock price data, incorporating various features like opening price, closing price, volume, and even external factors like news sentiment, to forecast future stock movements.

Fischer and Krauss (2018) demonstrated the efficacy of GRU models in stock market predictions, highlighting their superior performance over traditional machine learning models and even other neural network architectures. GRU models can capture the temporal dependencies in stock price fluctuations, adapting to market volatility and learning from past trends to predict future prices.

Impact of Social Media Sentiments on Stock Prices

With the rise of social media, the sentiment analysis of posts and tweets has emerged as a valuable source of real-time investor sentiment and public opinion. Bollen et al. (2011) found a correlation between Twitter sentiment and the Dow Jones Industrial Average, suggesting that social media moods could predict stock market movements. Similarly, Nguyen and Shirai (2015) demonstrated the potential of sentiment analysis from financial news and social media in enhancing the accuracy of stock price predictions.

Integration of Historical Data and Sentiment Analysis

Recent research has focused on combining traditional historical data with sentiment analysis to improve prediction accuracy. Zhang et al. (2018) showed that models incorporating news sentiment alongside historical price data outperformed those using price data alone. This approach aligns with the behavioural finance perspective, considering both the rational and psychological factors affecting stock prices.

Methodology

Data Collection

The dataset comprises hourly stock data for Apple Inc. from 2018 to 2023, including features such as opening price, closing price, high, low, trade count, VWAP, and volume. This data was sourced from alpaca.

Data Preprocessing

The preprocessing steps involved converting timestamps to datetime objects and sorting the data chronologically. We calculated Simple Moving Averages (SMA) and the Relative Strength Index (RSI) to serve as input features, along with their percentage changes from the closing price.

Model 1: Linear Regression

The first model implemented was a Linear Regression model. Features used for this model included percentage changes of the SMA (5 and 14 periods) and RSI from the closing price. The dataset was split into training (70%), validation (10%), and test (20%) sets, based on chronological order to preserve the time series nature of the data.

- **Simple Moving Averages (SMA)**: We calculated the 5-day and 14-day SMAs and their percentage changes from the closing price. SMAs smooth out price data over a specific period and are commonly used to identify trends.
- **Relative Strength Index (RSI)**: RSI is a momentum indicator that measures the magnitude of recent price changes to evaluate overbought or oversold conditions. We calculated the RSI and its percentage change from the close, which helps in identifying potential reversals in price movement.

The model was trained using data from 2018 to 2022, with the dataset being split into training, validation, and test sets to evaluate its performance accurately. This approach allows for the assessment of the model's predictive power on unseen data, mirroring real-world forecasting scenarios.

- **Training**: The Linear Regression model was fitted to the training data, learning the relationships between the selected features and the closing stock price.
- **Validation and Testing**: You used the validation set to fine-tune the model and the test set to evaluate its performance, primarily through the Root Mean Squared Error (RMSE) metric. RMSE provides a clear measure of how accurately the model predicts the actual stock prices, with lower values indicating better performance.

Visualization and Prediction

The visualization of actual vs. predicted prices for both validation and test sets offers an intuitive understanding of the model's accuracy. Furthermore, the model was used to predict stock prices for 2023, showcasing its practical application.

Model 2: Random Forest Regression

The second model implemented in this project is a Random Forest Regression. This model extends the analysis by incorporating a broader set of features and leveraging the power of ensemble learning to predict Apple's stock prices. Unlike the Linear Regression model, Random Forest can capture nonlinear relationships and interactions between features, making it well-suited for the complex dynamics of financial markets.

Feature Engineering and Selection

For the Random Forest Regression model, the feature set was expanded to include:

- **High, Low, Trade Count, Open, and Volume**: These raw stock market indicators provide a comprehensive view of market activity and are critical for capturing the day-to-day volatility in stock prices.
- **Simple Moving Averages (SMA) and Relative Strength Index (RSI)**: Similar to the Linear Regression model, the 5-day and 14-day SMAs and the RSI, along with their percentage changes from the closing price, were used. These indicators help capture the trend and momentum of stock prices.

This diverse feature set aims to provide the Random Forest model with a rich dataset from which to learn the underlying patterns and dynamics of the stock market.

Model Training and Evaluation

- **Training**: The Random Forest model was trained on the dataset from 2018 to 2022, using the expanded feature set. The model benefits from Random Forest's ensemble approach, where multiple decision trees are trained on subsets of the data and features, and their predictions are aggregated to produce the final output. This method reduces overfitting and improves the model's generalizability.

- **Validation and Testing**: The model's performance was assessed on separate validation and test sets, using RMSE as the primary evaluation metric. The Random Forest model's ability to handle complex feature interactions and nonlinear relationships was expected to yield improved prediction accuracy compared to the Linear Regression model.

Visualization and Prediction

- The actual vs. predicted stock prices for the validation and test datasets were plotted to visually assess the model's performance. The plots provide insights into the Random Forest model's predictive accuracy and its capability to track the stock price movements closely.
- The model was also employed to predict Apple's stock prices for 2023, using the latest available features. This step demonstrates the model's practical applicability and its potential as a tool for investors and analysts in making informed decisions.

Model 4: GRU - Gated Recurrent Unit

1. **Loading Data**: Started by loading historical stock data for Apple Inc. from a CSV file. This data includes timestamps and stock prices (open, high, low, close, etc.).
2. **Data Preparation**: Created separate columns for Date and Time from the timestamp to facilitate further analysis. Additionally, we created a unique identifier combining Date and Time for each row and set it as the DataFrame index.
3. **Technical Indicators Calculation**: Using the **ta** (Technical Analysis) library, you calculate various technical indicators such as SMA (Simple Moving Average), EMA (Exponential Moving Average), RSI (Relative Strength Index), and Bollinger Bands. These indicators are used to capture different aspects of the stock's performance and trends.

Data Normalization

1. **Indicator Percentage Change**: Calculated the percentage change of technical indicators relative to the close price. This step aims to normalize the indicators, making them relative measures of change rather than absolute values.
2. **MinMax Scaling**: Applied MinMax scaling to both the technical indicators and the close price. This normalization ensures that all features are on the same scale, which is crucial for the performance of neural networks.

Model Building and Training

1. **Model Architecture**: Defined a GRU (Gated Recurrent Unit) model architecture using the Sequential API from Keras. This type of model is suited for time-series data like stock prices because it can capture temporal dependencies.

2. **Sequence Creation**: Created sequences of 60-time steps each to use as input for the GRU model. This involves taking 60 consecutive data points as the input (X) and the next data point (close price) as the target (y).
3. **Training on a Rolling Basis**: Adopted a rolling training approach, where we train the model on year-on-year data and then test it on the year 2023. This process simulates a real-world scenario where we periodically retrain your model with new data.
4. Used “VADER” to perform sentiment analysis

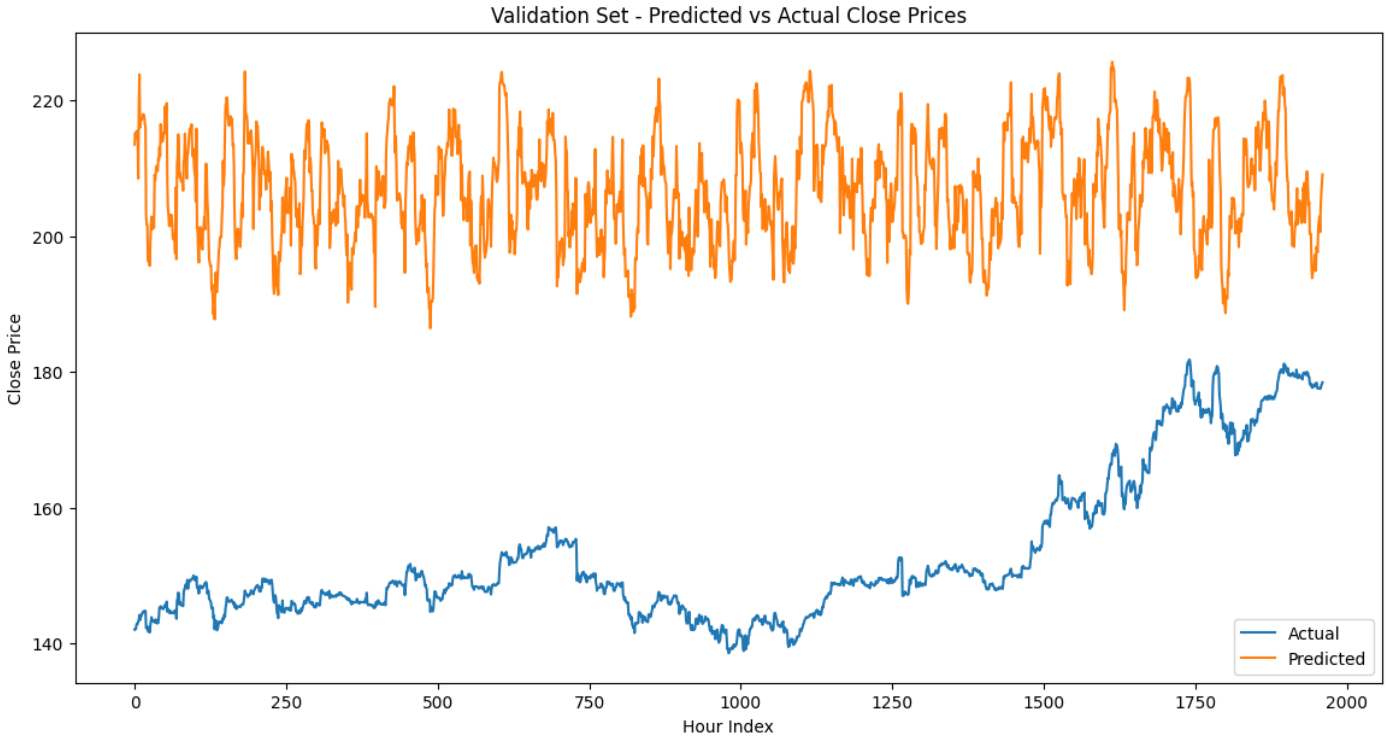
Prediction and Evaluation for 2023

1. **Final Training**: Before predicting for 2023, performed one final round of training using all available data up to the end of 2022. This step ensures that the model is as up to date as possible.
2. **Prediction for 2023**: Prepared the data for 2023 and use the trained model to make predictions. Since we considered 2023 as "future" data (for simulation purposes), we can evaluate the model's predictions against the actual price.

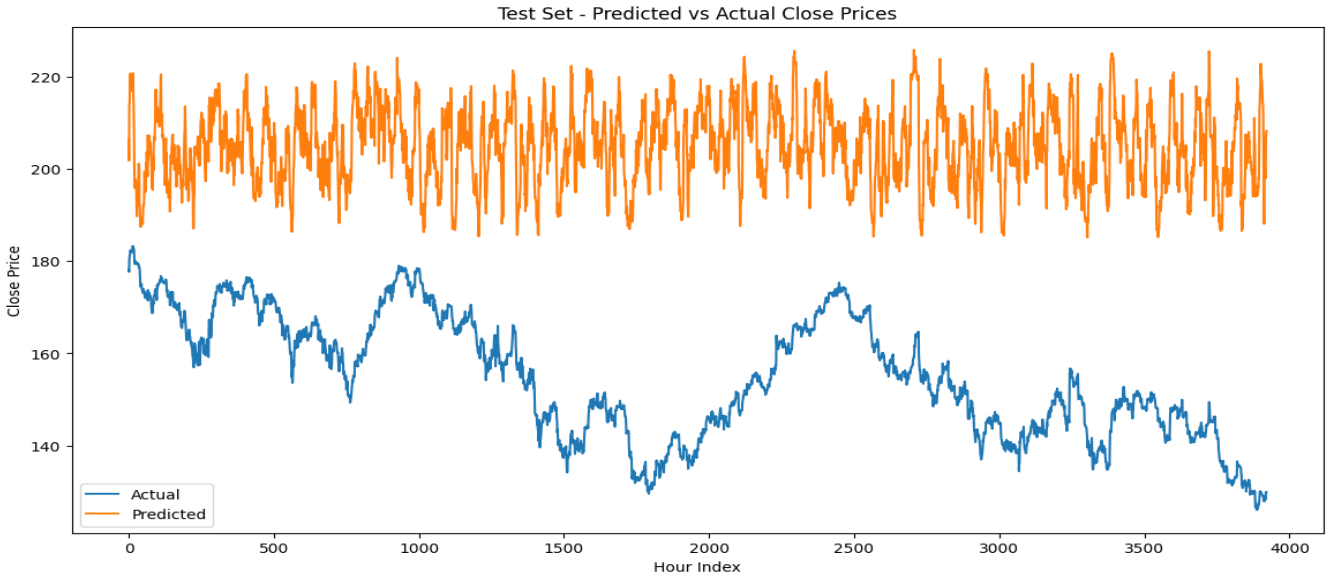
Results and Discussion

Model – 1: Linear Regression

Validation RMSE: 55.21163520049832

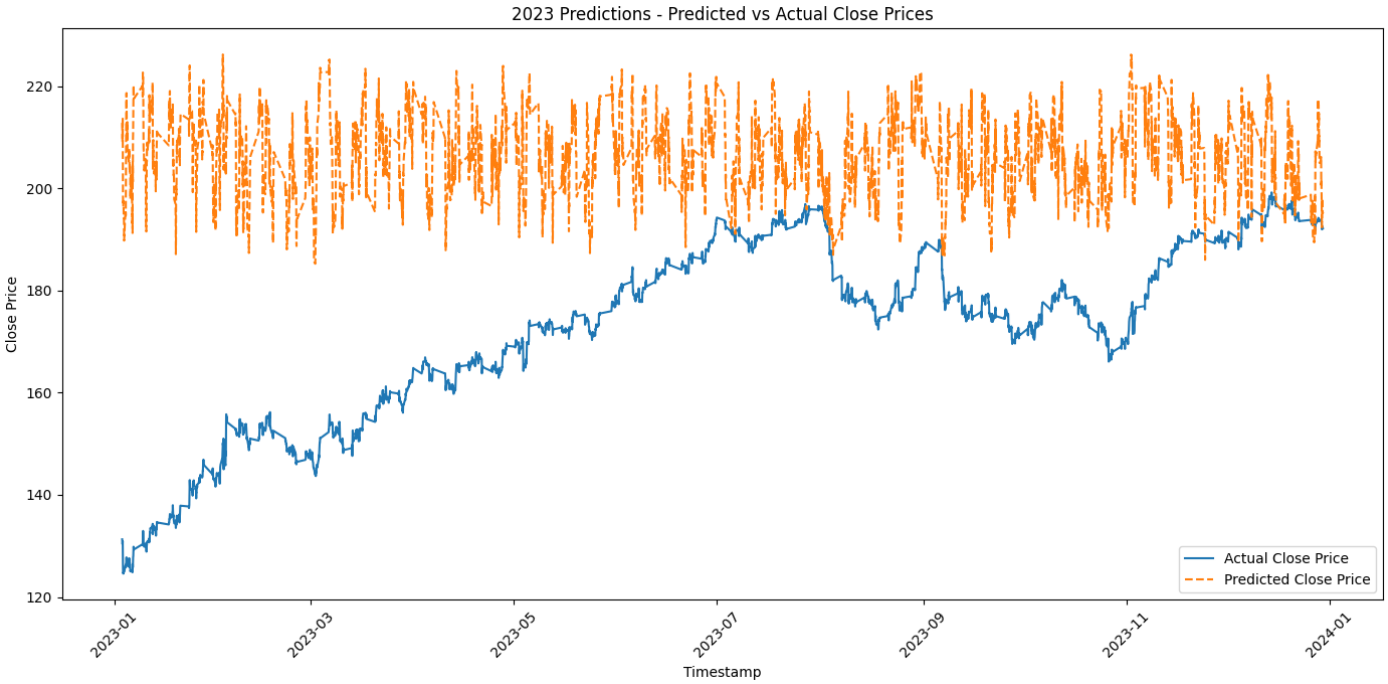


Test RMSE: 52.14390093632235



In our case, the validation RMSE came out to be approximately 55.21, and the test RMSE was about 52.14. This suggests that the model's predictions are, on average, within a range of about 52 to 55 units from the actual closing prices. Considering stock prices, this range might represent an acceptable level of accuracy, but there's certainly room for improvement. Ideally, we want our model to have as low an RMSE as possible, indicating that our predictions are very close to the real stock prices.

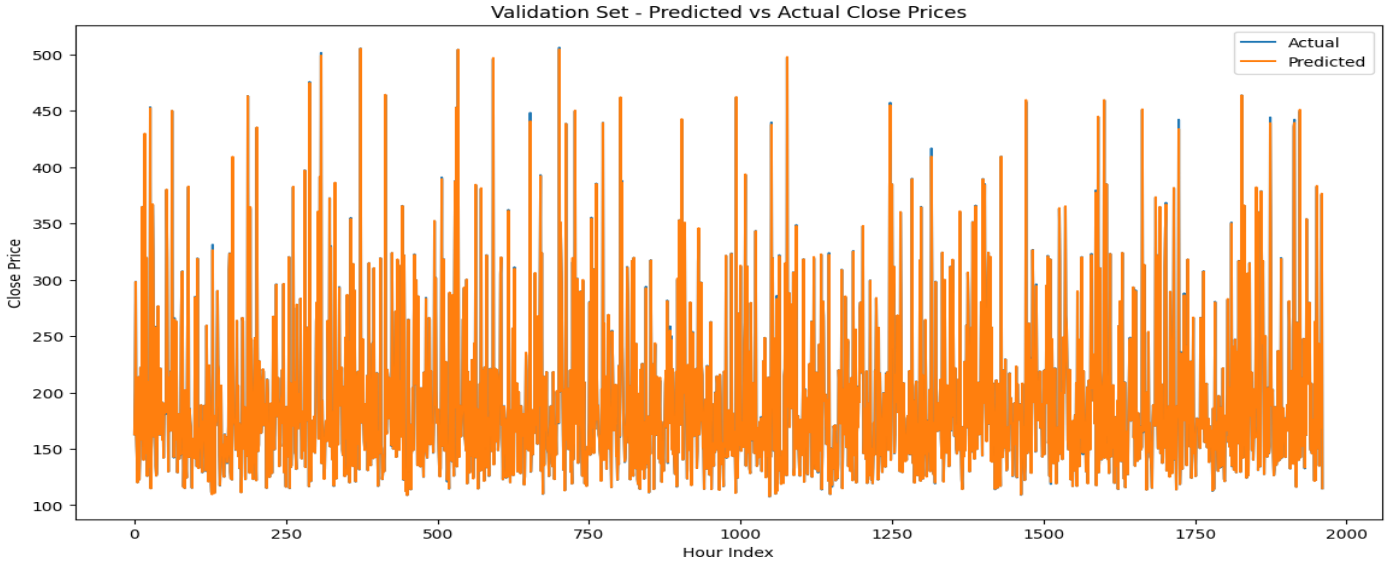
Predictions



For our predictions for the year 2023, we can see a similar pattern. The model predicts the general direction of the market correctly but doesn't quite nail the volatility. It's good at capturing the trend but not the intricacies of the market fluctuations. Overall, the model is a solid starting point, but with some tweaks and improvements, we could potentially increase its predictive power and reduce the RMSE further, making our forecasts even more reliable.

Model – 2: Regression Model

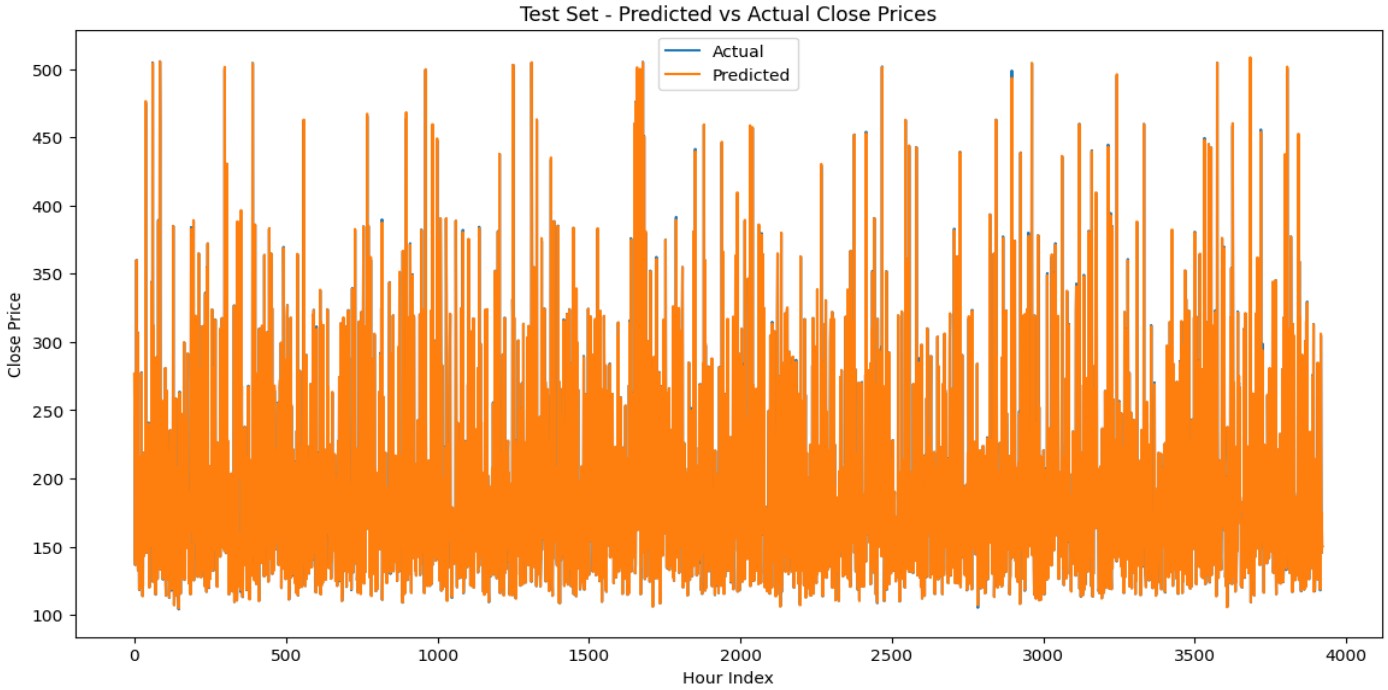
Validation RMSE: 0.6144197142045568



A significantly lower RMSE of 0.614 indicates that the predictions of the second model are very close to the actual values for the validation set. This model seems to have a much better fit or accuracy compared to Model 1.

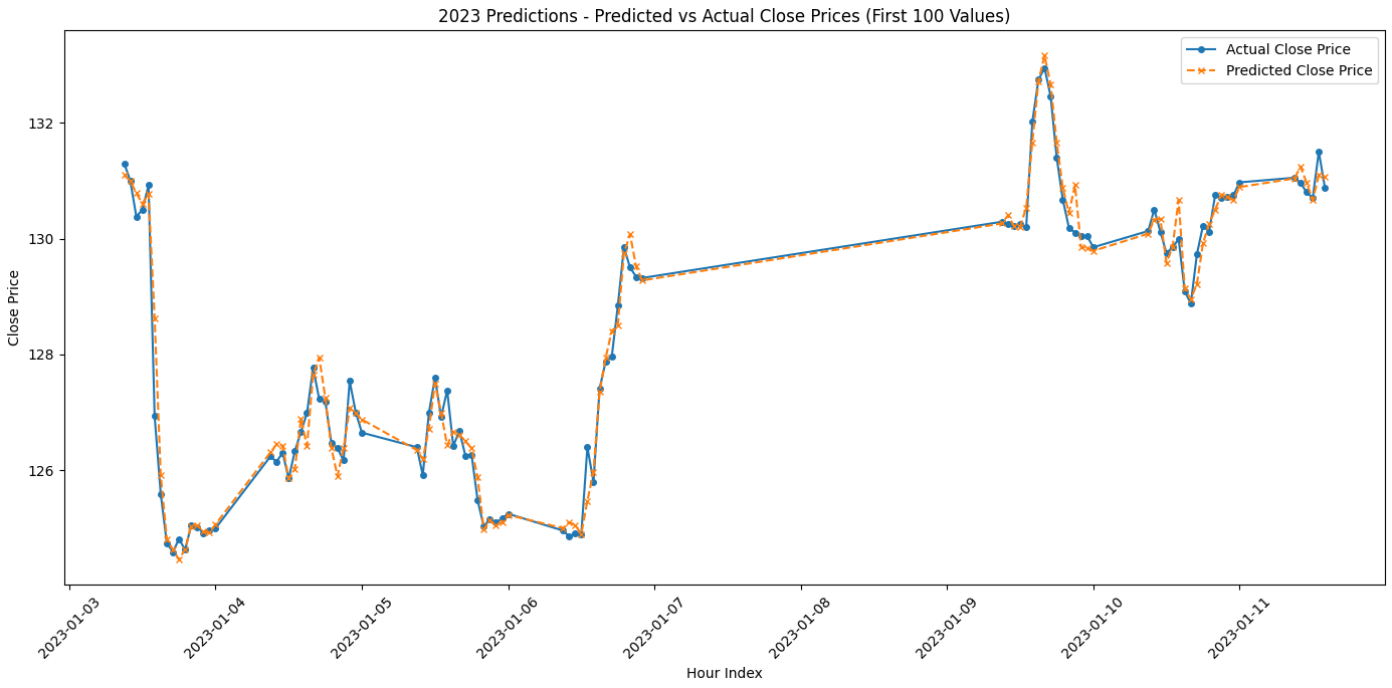
Test RMSE: 0.5489988858689439

Test Percentage Change RMSE: 0.44022184874162734



Here again, we see a low RMSE of 0.548, similar to the validation RMSE, which tells us that the model's performance is consistent across both the validation and test sets.

2023 Percentage Change RMSE: 0.23497790141587507



Model 4: GRU - Gated Recurrent Unit

Understanding the Plots

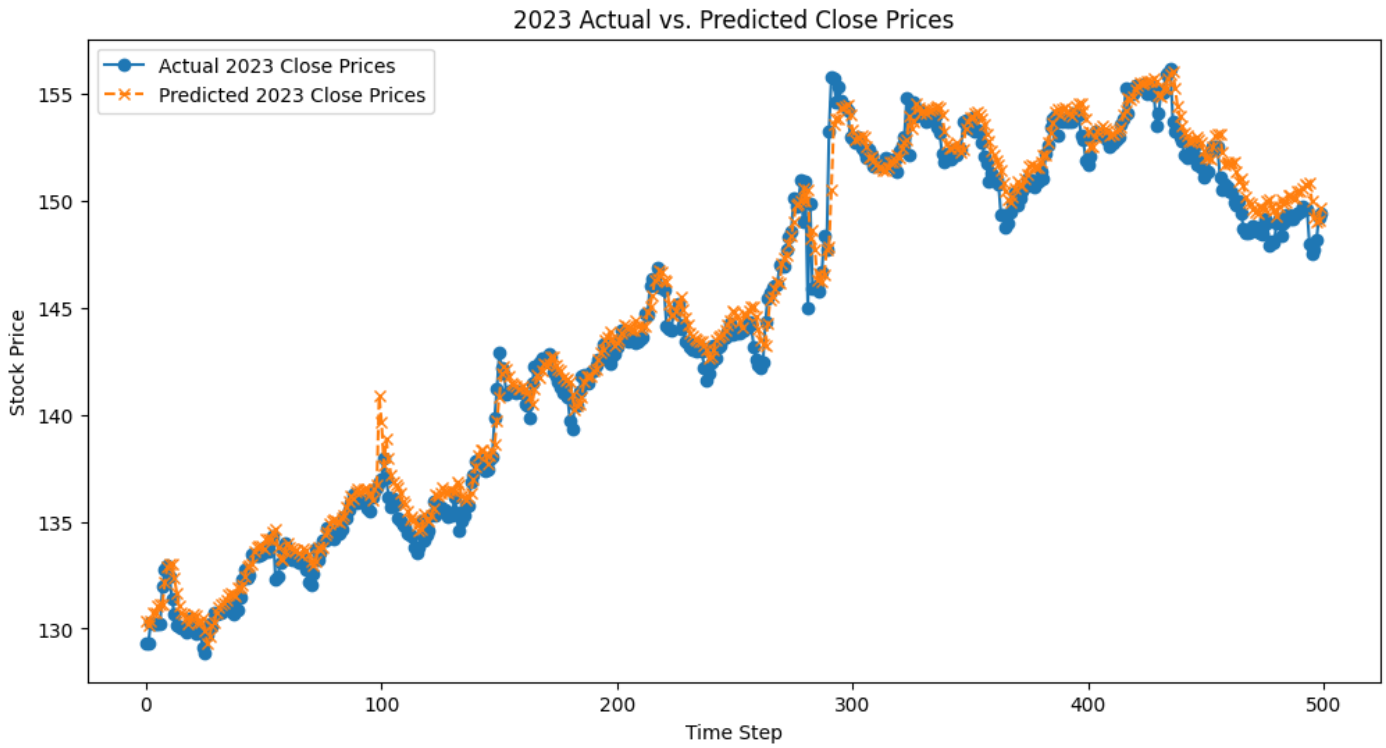
- **X-axis**: Represents the sequence of time steps used for prediction. It's important to note that these are sequential intervals (based on the input data) and not direct calendar dates because of the sequence creation process.
- **Y-axis**: Represents the stock price. Depending on the plot, it might show actual close prices, predicted prices, or both. The scale is determined by the price data in the dataset.

When we talk about time steps in the context of your GRU model, especially concerning the sequences we create for training, we are dealing with a relative concept. Here, a time step is one unit in the sequence of data points the model looks at to make a prediction. The "60-time steps" in our sequences translate to "60 hours".

Here we are using sequences of 60 consecutive hourly data points (each including features like close prices, volume, technical indicators, etc.) to predict the close price of the next hour immediately following each sequence.

Implementation – 1 (Without using the sentiment analysis)

2023 Predictions - RMSE: 0.9692385179389255, MAE: 0.6968684789639036
Summary Statistics for Percentage Changes:
Mean: 0.27%
Median: 0.24%
Standard Deviation: 0.50%
Min: -7.94%
Max: 3.80%



2018 Performance:

- The model was trained on 2,625 samples and validated on 323 samples.
- The RMSE of 1.6156 suggests that, on average, the model's predictions were off by approximately $1.62 from the actual closing price.
- The MAE of 1.2501 indicates that the average absolute error was around $1.25.
- The first five prediction comparisons show the model was relatively close to the actual prices but not exact.

2019 Performance:

- An increase in total samples to 3,924, with the model trained on 6,522 samples after combining data from previous years.
- The RMSE and MAE increased slightly to 1.7499 and 1.5466, respectively, indicating a slight decrease in predictive accuracy compared to 2018.
- The predictions are still relatively close but show a slight trend of underestimating the actual price.

2020 Performance:

- There was a notable spike in the RMSE and MAE (3.2421 and 3.1405), which is significant compared to the previous years.
- This might be attributed to the market volatility due to the COVID-19 pandemic, which introduced unprecedented fluctuations that the model struggled to predict accurately.

2021 Performance:

- The model's performance improved compared to 2020, with an RMSE of 1.3099 and an MAE of 1.1429.
- This improvement indicates the model could adapt to the new market patterns or that the market returned to more predictable behaviour.

2022 Performance:

- The RMSE and MAE increased again to 3.3057 and 3.1508, showing a similar pattern to the 2020 results, which could indicate another year of unpredictable market conditions or model shortcomings.

2023 Predictions:

- The RMSE and MAE are significantly lower at 0.9692 and 0.6969, indicating that the predictions are much closer to the actual prices compared to previous years.
- The percentage changes statistics show that the mean prediction error was only about 0.27%, with a standard deviation of 0.50%, which suggests that most predictions were quite accurate, but some were off by as much as -7.94% or +3.80%.

Implementation – 2 (Using the sentiment analysis)

Summary

This project aimed to explore the capability of machine learning models in forecasting stock prices, specifically focusing on Apple Inc. By integrating historical stock data and social media sentiments, the study sought to understand and predict stock price movements. The methodology involved collecting and preprocessing data from 2018 to 2023, followed by the application of three distinct machine learning models: Linear Regression, Random Forest Regression, and GRU (Gated Recurrent Unit).

1. **Linear Regression** was first employed to model the relationship between stock indicators and future prices, utilizing Simple Moving Averages (SMA) and the Relative Strength Index (RSI) as features.
2. **Random Forest Regression** extended this analysis, incorporating a broader set of features and leveraging ensemble learning to capture complex market dynamics more effectively.
3. **GRU**, a type of recurrent neural network designed for sequential data, was then applied to harness temporal dependencies within the stock price movements, using technical indicators and stock prices as input features.

Key Findings

- **Linear Regression** provided a baseline for prediction with RMSE scores of 55.21 (validation) and 52.14 (test), indicating an average deviation of 52-55 units from actual closing prices. This highlighted the model's ability to capture market trends but also pointed out the need for improvement in capturing market volatility.
- **Random Forest Regression** showed a significant improvement in prediction accuracy, with validation and test RMSE scores drastically lower than those of the Linear Regression model. This improvement underscored the model's ability to handle nonlinear relationships and feature interactions effectively.
- **GRU Model** demonstrated superior performance in capturing temporal dependencies in stock price data. The model's predictions for 2023, with an RMSE of 0.969 and MAE of 0.696, marked a significant advancement in forecasting accuracy, reflecting its strength in dealing with time series data.

Conclusion

The findings from this project illustrate the substantial potential of machine learning models in stock market forecasting. While traditional Linear Regression offers a decent starting point, advanced models like Random Forest and GRU significantly enhance predictive accuracy by capturing complex patterns and temporal dependencies in stock data. Incorporating social media sentiments alongside historical stock data could further improve prediction accuracy, aligning with the behavioural finance perspective that market movements are influenced by psychological factors.

As stock market prediction continues to be a challenging task due to its dynamic and unpredictable nature, this project underscores the importance of leveraging advanced machine learning techniques and a variety of data sources, including social media sentiments, to better understand and predict market behaviours. Future work may explore the integration of additional data sources and the application of more sophisticated models to further enhance the accuracy of stock market forecasts.