Stock Market Prediction Analysis

StockPredictor

This document explains the methodologies used for predicting stock price movements using advanced machine learning models. The analysis involves data from major stock tickers, applying Random Forest and Long Short-Term Memory (LSTM) networks to forecast significant price movements and future stock prices, respectively.

Data Collection

Data for various stock tickers is collected using the Yahoo Finance API. This dataset includes historical prices from January 2021 to the present, capturing daily opening, closing, high, low prices, and volume for each stock. This comprehensive dataset provides the foundation for our subsequent analyses.

Data Preprocessing

Percentage Change Calculation

Daily percentage changes in the adjusted close prices are calculated to understand day-to-day volatility. A significant daily percentage change is defined as a movement greater than 5%. This threshold helps in identifying days with unusually high volatility, which are critical to predicting future price movements.

Lag Features

Lag features are created by shifting the significant movement indicator for five days. These features help in capturing temporal dependencies, which are essential for time-series forecasting in financial markets. They provide the model with information about how significant movements in stock prices were preceded by similar movements in the recent past.

Feature Engineering

Lagged features of significant movements are used as predictors. This approach is based on the hypothesis that past volatility patterns can provide insights into future price behaviors, which is a common assumption in financial time series analysis.

Model Training and Evaluation

Random Forest Model

A Random Forest Classifier is trained to predict whether there will be a significant price movement on a given day. This model was chosen for its effectiveness in handling non-linear relationships and its robustness to overfitting with high-dimensional data.

Evaluation Metrics The F1 score is used to assess the model's performance, focusing on the balance between precision and recall. This metric is particularly useful in the context of imbalanced datasets, like ours, where significant movements are rare events.

LSTM Network for Price Prediction

Model Setup   An LSTM network is used to predict future stock prices based on historical data. This type of neural network is particularly suited for sequential data like stock prices, where the order of data points is crucial.

Scaling   Data is scaled using MinMaxScaler to normalize input features, ensuring that the LSTM model receives input data within a scaled range that optimizes internal weight updates during training.

Training   The model is trained on historical price data, learning to predict the next day's closing price based on the patterns observed in the past 100 days. Training involves adjusting the model weights to minimize prediction errors, measured by the mean squared error loss function.