

Stat4012 Group-Project

Overview

To synthesize profitable trading strategies, accurately predicting future stock prices is crucial. There are mainly three approaches to make such predictions: cross-sectional regressions, traditional time series models and machine learning models.

Since machine learning models have shown higher accuracy even for volatile data, we hope to find an effective stock price forecasting model by comparing the prediction accuracy of two deep learning algorithms: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). We will also compare the performance of our selected model with the Random Forest algorithm to test its capability.

Project Design

Data Collection

Stock price and relative information will be download from Bloomberg terminal or [Yahoo Finance](#). A single stock (i.e. Tencent 0700.HK) historical pricing data (`High` , `Low` , `Open` , `Close`) and trading `Volume` at 5-minute intervals in Hong Kong stock market for past 3 year will be collected. `High` , `Low` , `Close` and `Volume` will be the regressors (features) while `Open` price will be the response.

Data Preprocessing

Data will be preprocessed by considering the missing data, outliers, stock split and dividend. Returns of the stock will be calculated and converted into a stationary time series sequence. Then it will be divided into training set and testing set with rate 8:2 for further analysis.

Neural Network Development

Our neural network model will start from univariate CNN with the return be the feature. Multivariate CNN model will be developed by combining features such as `Volume` after change the data shape. We also want to build a univariate LSTM model based on with previous returns as input to predict the `Open` price for the next few days.

For each model, we will try different hyperparameters to find the model with the best performance, including number of layers, number of neurons in each layer, batch size, activation function, optimizer. Unnecessary nodes will also be pruned to improve the models.

Model Training

Train the models by backpropagation to minimize the loss function.

Model Comparison

Models using algorithms CNN, LSTM and Random Forest will be tested on the testing set respectively. We will conduct hypothesis tests following Diebold and Mariano (1995) and Giacomini and White (2006) to examine the relative predictive ability of different models. Accuracy, precision, recall, F1 score, MSE and ROC will be used to measure the performance of each model.

References

1. [Analysis and Forecasting of Financial Time Series Using CNN and LSTM-Based Deep Learning.](#)
2. [Integrating Fundamental and Technical Analysis of Stock Market through Multi-layer Perceptron.](#)
3. [How to choose the number of hidden layers and nodes in a feedforward neural network?](#) ¹
4. [Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators](#)

1. How to decide the number of layers and neurons in each layer?: Refer to Reference 3. [↗](#)