Machine Learning Engineer Nanodegree

Investment and Trading Capstone Project Proposal

Build a Stock Price Indicator

Yipeng Shi
Oct 26st, 2017

Proposal

Domain Background

In today's competitive markets, managers are hoping to discover new predictive patterns in data, through technologies like artificial intelligence, to gain an edge over rivals. [1]Machine learning is an application of artificial intelligence (AI) that allows computer systems to use algorithms to adapt to enable better outcomes, based on data rather than explicit programming. [2]It excels at finding patterns and making predictions, and used to be the preserve of technology firms. The financial industry has started to make use of it. One of the earliest investors in the industry even forecasted that machine learning poses a threat to equity hedge funds within the next decade as the technique becomes powerful enough to forecast market moves better than humans.[3]

Machine-learning has found its application in many different areas of finance. It has been applied to spot unusual patterns of transactions, which can indicate fraud. It has been applied in the document-heavy parts of finance, where AI-based systems can be used to recognize text. It is also good at automating financial decisions, like assessing creditworthiness or eligibility for an insurance policy. The newest frontier for machine-learning is in trading, where it is used both to analyze market data and to select and trade portfolios of securities. [4] This project will utilize machine-learning in this area to build a stock price indicator.

[1] Segal, J. 2017, "Quants Skeptical of Big Data, Machine Learning Hype", *Institutional Investor* .

[2] *First Derivatives: Major Machine Learning Investment in Kx Technology* 2017, Melbourne.

[3] Fortado, L. & Wigglesworth, R. 2017, "Machine learning set to shake up equity hedge funds", *FT.com* .

[4] *Unshackled algorithms; Machine-learning in finance* 2017, , The Economist Intelligence Unit N.A., Incorporated, London.

Problem Statement

The problem to be solved in this project is to accurately predict the future closing value of a given stock across a given period of time in the future. One potential solution is to utilize supervised learning methods such as linear regression based on a time series of adjusted closing stock price values to predict the future value. This model can be trained using the historical data of different stocks, and the predicted value can be compared with the actual value of these stocks. The percentage difference between the predicted value and the actual value can be used to calculate the accuracy of the model.

Datasets and Inputs

This project will use open source historical stock price data from Yahoo! Finance. Yahoo finance provides a relatively simple process of obtaining basic financial information. Information available includes financial statements and price and volume data. CSV files can be obtained through its website about the data, and web API is available.

Although data from Yahoo is free, the accuracy of its information is reliable.[5] Actually, a study by (Flanegin, et al 2009) suggested that Yahoo Finance data was acceptable for research purposes. The dataset will be imported and processed in the project. It will be split into training and testing set. The features in the dataset will be used to train the models and make prediction.

Solution Statement

The solution to the problem is the predicted closing price across a given time in the future, based on the input data of a certain period of data for a certain stock. A good machine learning model should be able to make relatively accurate prediction for the future price of a given stock. The percentage difference between the predicted price and the actual price can be used as the metric of how well the model works.

Benchmark Model

[5] Bailey, B.A., Dennick-Ream, Z. & Flanegin, F.R. 2014, "CREATING AN AACSB TECHNOLOGY CLASS FOR FINANCE MAJORS UTILIZING BLOOMBERG, EDGAR, YAHOO FINANCE, AND MICROSOFT EXCEL", *ASBBS E - Journal,* vol. 10, no. 1, pp. 74-82.

Stock market indexes are usually used as the benchmark for stock market prediction.[6] In this project, I will use S&P 500 index as the benchmark. In this benchmark model, the price of a given stock will be calculated assuming the stock moves the same way as the S&P 500 index does. The predicted closing price across a period of time in the future is calculated as the product of the closing price of the last day of the input date and the percentage change of S&P 500 index across this period of time. This predicted price from the benchmark model will be compared with the predicted price from the machine learning model.

Evaluation Metrics

The measures of performance for this project will be the percentage difference between predicted and actual closing values of the target stock. For more input of more than one stock, the mean percentage difference will be used as the metric.

Project Design

The project will be implemented with Python notebook and the corresponding python machine learning packages. It will include the following different stages.

1. Set Up Infrastructure

   Python notebook and corresponding Python packages such as *sklearn* will be imported

2. Data preparation and exploration

   Stock data will be imported from csv, and the data will be explored by calculating statistics and making plots. The data will also be split into training model and testing model.

3. Develop supervised machine learning model

   Models such as linear regression and KNN will be developed to fit the training data. The model parameters will be adjusted to avoid overfitting and underfitting.

4. Model evaluation

[6] Anon, (2017). [online] Available at: What data is typically used to benchmark a stock market prediction task? (n.d.). Retrieved October 29, 2017, from https://www.quora.com/What-data-is-typically-used-to-benchmark-a-stock-market-prediction-task [Accessed 29 Oct. 2017].

The model will be tested against testing data. The result will be measured against the evaluation metrics and be compared with the benchmark model.

5. Discussion and conclusion

   The result will be discussed, and the conclusion will be recorded into the project report.