Proposal: Investment and Trading Capstone Project
Build a Stock Price Predictor

March 20, 2017

Soyoung Park

1. Domain background

Technical stock price analysis can be approached as a typical time series data analysis without understanding of health of a particular security while fundamental analysis requires deep domain knowledge about a security or domain of interest. In this project, one particular security, Intel stock, will be chosen for forecasting the return of 7,14,28,120 days, and its return on invest will be backtested for the decisions of trading based on the increase/decrease classification prediction.

2. Problem statement

Many day stock traders use their own various strategies to find patterns in one particular stock price chart and make their bets without having to understand the health of business or industry because they believe all information is already baked in the stock price. [1] This type of technical analysis process leading to investment is not only risky due to lack of fundamental analysis or statistical significance, but also, takes long time to build experience to make a consistently reasonable performance. If technical analysis is believed as valuable, machine learning/deep learning can be very good alternative solution to guide the day traders to take a conventional historical chart and help to understand the trend of price changes more systematically along with the accuracy. This can help the day traders to have more confidence in decision making of trading whenever a signal is shown because model accuracy is always considered for any prediction.

3. Datasets and inputs

Three approaches are planned to predict the stock trend by:

(1) regression of adj.close price itself after 7,14,28,120,365 days,
(2) regression of return of 7,14,28,120,365 days,
(3) Multi-class classification of return (or log regurn) of 7,14,28,120,365 days
    (a) Gain(class 1): equal to or greater than +10%
    (b) No change(class 0):-10<x<+10%
    (c) Loss(class -1): equal to or greater than -10%

Daily adj. close price and volume data (5080 points) from January,1,1997-March,9,2017 will be pulled from Yahoo! Finance [2]. Precisely,

3 different targets will be used to predict for different prediction approaches:

"INTC_price"

or

"Return of 7,14,28,120,365 days"
"Log return of 7,14,28,120,365 days"

Or,

Category of increase/no change/decrease of adj.close price.

by using the following 13 technical features:

"SPY_price": S&P 500 daily adj. Close price

"SMA20": simple moving average of adj. Close price in 20 days
"SMA50":simple moving average of 50 days
"SMA200": simple moving average of 200 days

"Rolling_mean": same as "SMA20"
"Rolling_std":standard deviation of price in 20 days

"Bollinger_upperband": Rolling_mean+2*Rolling_std
"Bollinger_lowerband":Rolling_mean-2*Rolling_std

"EMA_slow": exponential moving average of 26 days
"EMA_fast": exponential moving average of 12 days

"RSI": relative strength indicator. [3]
"MACD": Moving Average Convergence and Divergence. EMA_fast-EMA_slow

The dataset will be split into 3 (for training) to 1 (for testing) ratio for the given the time-series nature of the problem to avoid any peaking future data into training data as following:

- Shape of training set : features (3810, 13) and target (3810,) from **1997-01-02 to 2012-02-21**
- Shape of testset :feature (1270, 13) and target (1270,) from **2012-02-22 to 2017-03-09**

- **Query set can be anytime after 2012-02-21.**

Any missing data from the original set will be either backward filled(before the very first datapoint) and forward filled(in between any rest of data) and the initial return before 7,14,28,120,365 days will be filled with zero.

4. **Solution statement**

The ultimate solution of this project is to give buy/sell/hold decision based on classification task and investment will be made automatically. Initial capital is $100,000 and total 50 shares of an asset will be transactioned at the opening price depending upon the direction of the forecast, and the closing total price of the day will be calculated. And this can be compared with the price prediction by regression tasks assuming the trading has not happened. For regression, SVR, random forest, and adaboost,Knn regressor, DecisionTree regressor will be used to predict the price or return of adjusted close price after 7,14,28,120,365 days. For multi-class classification, SVM, random forest, and adaboost,Knn, DecisionTree classifiers will be used. Also, feature importance will be studied for the cases that are trained with random forest and adaboost [4][5]. Once key features are determined and selected, the dataset will be re-trained to improve accuracy.[6]

5. **Benchmark model**

Simple knn regression on the adj. close price will be a baseline. This will be compared against in terms of (1) prediction RMSE with respect to other models, and (2) total return on invest in percent over the course of investment period by trading/backtesting.

6. **Evaluation metrics**

RMSE of price/return/log-return and accuracy of classification between target of test set and prediction will be used as evaluation metrics of each approach/model.

7. **Project design**

A training interface will accept two inputs:

- Ticker symbol of a security of interest :”INTC”
- Time range (start_date= “January,1,1997”, end_date=”March,9,2017”)

The stock behavior will be modeled based on several regressor/classifiers including knn, SVM, random forest, and adaboost,etc.

A query interface will not be required since there will be a plot of total price return of initial capital $100,000 over the period of time after training vs. simple regression, but it can accept

- Investment period after the end of training date (7,14,28,120,365 days)

to read a single number of return as output for the total price value result by backtest. The backtest methodology is referred from [7].

A basic run of the core system would involve one call to the training interface, (and one call to the query interface only if desired).

At the end, there will be a discussion about whether which technical indicator is most influential, which model, between regression and classification, would give the highest performance in the report.

8. References

[1]http://www.investopedia.com/trading/how-to-become-day-trader/
[2] pandas.io.data.get_data_yahoo(ticker symbol, start_date, end_date) will pull daily price, volume data for the given ticker symbol, period of time.
[3]https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/RSI
[4]https://www.quantopian.com/posts/machine-learning-on-quantopian

[5]http://cs229.stanford.edu/proj2013/Lin-Feature%20Investigation%20for%20Stock%20Market%20Prediction.pdf

[6]http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf

[7] #http://francescopochetti.com/stock-market-prediction-part-introduction/