Machine Learning Engineer Nanodegree

Capstone Proposal

Valentyn Masyakin
November 10th, 2017

Domain background

Stock market business is by far the fastest investment of all around the world. Provided you know little bit of economics, understand the opportunities and risks, have enough idle savings to invest, investing in stocks can yield high profits. There are many trading strategies developed up to date – Day Trading, Position Trading, Swing Trading etc., but in the real life, stock market resembles a living organism that constantly evolves – that can work today and may fail to work tomorrow. Implementing flexible strategy that is able to observe the environment and react to changing condition quickly versus having fixed strategy is important for the successful trading. . The strategy relies on the application of machine learning to predict prices. Machine learning if widely used for such tasks - some examples of highly-reputable firms that do this include Two Sigma Investments, D. E. Shaw (company), Renaissance Technologies (hedge fund), Hudson River Trading etc.. These companies are consistently successful in these automated trading strategies, generating very high returns for their clients (Quora). One of the research of the application of machine learning to calculate future prices "A Machine Learning Model for Stock Market Prediction" is published by Osman Hegazy, Omar S. Soliman and Mustafa Abdul Salam.

Problem Statement

According to the Efficient Market Hypothesis (Efficient Capital Markets: a Review of Theory and Empirical Work), developed by Eugene Fama, markets are rational, and all available information is adequately reflected in the stock prices. Even in its weak form, it says that one cannot use the past prices data to predict future prices. Considering this hypothesis, it seems to be impossible to use the technical analysis to predict future prices because market reacts rapidly and any current price has already absorbed the news. However many psychological researches suggest that people tend to underreact to the news in the short term and overreact to them in the long term (Combs). We will attempt to exploit this fact and use the technical analysis of stock market to predict future 'Adj Close' prices of the selected individual stocks based on historical stock market performance (Open price, Close price etc.) and build a portfolio in this project.

In finance, a portfolio is a collection of investments held by an investment company, hedge fund, financial institution or individual. It is important to diversify the portfolio in order to reduce risks by allocating investments among various financial instruments, industries and other categories. We create portfolio assets allocation by taking into account expected returns based on the predicted prices and associated risks (prediction confidence).

Datasets and Inputs

In this project, we will be using publicly available S&P500 companies' historical stock prices dataset that can be obtained from Yahoo! Finance web-service. Machine learning is known to be very efficient finding patterns in consistent data, however, it cannot predict artificial events that drastically affect markets such as global crisis, war etc. Thus, we shall exclude stock price data contaminated by world economic crisis of 2008. The chosen dataset contains aggregated information about daily transactions since 2010-01-01 and has the following structure:

Field	Description
Ticker	Company ticker that uniquely identifies business entity
Industry	Company industry
Date	Operation date
Open	Open price for the target day
High	Highest price among the daily transactions
Low	Lowest price among the daily transactions
Close	Closing price for the target day
Volume	Number of stock sold
Adj Close	**(Outcome)** Closing price that reflects price anomalies e.g. Stock splits, Dividends

In this way, the number of data points per company will be around 1700, which might not be enough for efficient learning. We will apply roll forward cross validation technique to increase the number of data points and avoid "peeking into the future" situation when training set follows testing.

Other input parameters includes:

- Date – date forecast starts
- Training period, days – number of days to train;
- Forecasting period, days – number of days to predict;
- Initial portfolio allocation – list of weights of the initial default portfolio;
- List of companies – companies that used for the training.

Output data and parameters:

- The outcome of machine learning is 'Adj Close' field of the dataset as it reflects price anomalies
- Portfolio – dictionary of {ticker: allocation weight} that identifies proposed portfolio allocations
- Portfolio statistics – portfolio performance information

Solution Statement

As described in the **Problem Statement**, we will be using supervised regression learning to predict future prices for the given period. The next step is to build a portfolio allocation for different combination of tickers and identify the one that provides best performance.

We will be using different supervised learning regression algorithms to predict future prices – Linear regression, Support vector machines, Random trees, Ensemble methods etc. Source data must be pre-processed before passing it to the learning algorithms. Pre-processing techniques to be used are: data normalization, converting absolute values to its changes, handling missing data, one-hot encoding to pivot categorical features.

The application workflow is as follows:

1. Future price prediction:
   1.1. Read stock market data;
   1.2. Add engineered features;
   1.3. Massage and pre-process data to eliminate and/or fix anomalies;
   1.4. Train the model according to given period;
   1.5. Predict prices using trained model;
2. Build portfolio based on the obtained future prices:
   2.1. Calculate portfolio statistics for different combination of allocation;
   2.2. Choose the best combination based on expected returns and risks. We will use Sharpe ratio to identify the best performing portfolio.

Benchmark Model

To assess the performance of our predicting model we will compare it with the results of a default version of the random forests algorithm. We also compare it over the baseline exponential weighted moving average (EWMA) function to determine the minimum bound of predictive power of our model.

We also use backtesting of the model. Backtesting – the process of testing a trading strategy on relevant historical data to ensure its viability. We will simulate the trading of a strategy over an appropriate period and analyze the results for the levels of profitability and risk.

Portfolio statistics information with the metrics below will be calculated to assess its performance

- daily returns;
- cumulative returns;
- avg daily returns;
- std deviation of daily returns;
- Sharpe ratio;

Evaluation metrics

In order to evaluate the prediction algorithm we will be using the following metrics:

1. RMS (root mean squared error) – to assess the deviation of predicted prices from the real ones on test dataset;
2. Correlation of actual and predicted prices.

Project Design

The solution proposed includes two modules: predicting future price and build portfolio allocation.

Predicting future prices

In the first module, we will attempt to predict future stock prices using supervised learning. We will download the historical prices data from publicly available Yahoo! Finance source using Python library fix_yahoo_finance for all companies in the S&P500 list. To eliminate unnecessary calls to API, we will save the data in the local folder after the first run, and then refresh data incrementally during the following executions.

There are countless numbers of predictive factors that can be derived from the initial data

source and choosing which ones to use is not a trivial task. We will attempt to engineer the features that are important for the predicting including the following:

- Simple moving average for short and long periods;
- Exponential moving average for short and long periods;
- Price rate of change;
- Bollinger bands – describes deviation amplitude;
- SPY data for the corresponding date is important to estimate overall market performance
- Day of week.

After preliminary testing of used learning algorithms, it might be necessary to alter the list of features.

It is important to shape in and order the data to make it convenient for the learning consumption. The important information for the learner lies not in real values, but in its change from on point to another, so we will modify our dataset to keep relative values rather than absolute. In order to optimize the machine learning process, the data should be normalized. We choose Z-score method as it is one of the most common ways of data normalizing. The formula is as follows:

$$\mathbf{z = (x - \mu) / \sigma}$$

where:

Z – the standard score

X – real value

M – the mean and σ is the std

We will pack above information about each day in training period, starting current date, sequentially to generate prediction for i-th future day.

Date	X
Feature set 1	
Feature set 2	
…	
Feature set N	
Predicted price day index	
Future price	Label

where:

N – numbers of days in training;

Predicted price day index – in the range [1, Forecasting period];

Feature set - selected set of features for each ticker in the provided List of companies.

Different regression algorithms will be used to predict target future prices:

- Linear regression;
- SVM;
- Ensemble methods
  - Adaboost
  - Random regression forests
  - Gradient boosting

Build portfolio

We will identify the best performing stocks and the build portfolio according to input parameters with predicted prices. We will be using Sharpe ratio as a criteria to allocate weights. The

Sharpe ratio is the average return earned in excess of the risk-free rate per unit of volatility or total risk and can be calculated with the formula:

$$= \frac{\bar{r}_p - r_f}{\sigma_p}$$

Where:

$\bar{r}_p$ = Expected porfolio return

r_f = Risk free rate

σ_p = Portfolio standard deviation

References

Combs, A. T. (n.d.). *Python Machine Learning Blueprints: Intuitive data projects you can relate to.*

Fama, E. F. (n.d.). Efficient Capital Markets: a Review of Theory and Empirical Wor.