Machine Learning Engineer Nanodegree

Capstone Proposal

by Jason Carter

November 15, 2016

Proposal

Investment and Trading - A Stock Price Indicator

Domain Background

An investment is an asset or item that is purchased with the hope that it will generate income or will appreciate in the future. [...] In finance, an investment is a monetary asset purchased with the idea that the asset will provide income in the future or will be sold at a higher price for a profit.

This proposal will target the finance domain, specifically, the subset of stocks and the stock market. Going forward, the definition of investing and trading will be accepted as:

- **Investing** is the act of committing money or capital to an endeavor (a business, project, real estate, etc.) with the expectation of obtaining an additional income or profit.
- **Trading** is a basic economic concept involving the buying and selling of goods and services, with compensation paid by a buyer to a seller, or the exchange of goods or services between parties. [...] In financial markets, trading refers to the buying and selling of securities, such as the purchase of stock on the floor of the New York Stock Exchange (NYSE).

The First Stock Market

The world's first stock markets are generally linked back to Antwerp, Belgium as far back as 1531. But it wasn't until 1602, when the Dutch East India Company officially became the world's first publicly traded company.

Today, tens of thousands of companies are publicly traded on the stock market around the world with millions of investors, ranging from casual individuals and professional day traders to high speed volume traders and large hedge fund firms.

In Search of Profit, knowledge and Understanding

From 1602 'til now, the ultimate goal of any investor has been to make profit. Being able to **predict** which assets will appreciate or depreciate in value over time, is obviously advantageous to the owner of said knowledge. Whether it be individual investors, hedge fund firms or academic researchers there has always been interest in the stock market in search of profit, knowledge and Understanding.

Personal Motivation

The world of machine learning permeates many domains, and on a personal level finance happens to be one of the more interesting areas. Machine learning can not only help solve complicated financial problems but also can help the average citizen grow their personal wealth (directly or indirectly) whom otherwise would not have the opportunity due to income or social level. Whether this be via analysis of their financial spending and budgeting or having access to financial tool such as "roboadvisors", this data and the application of machine learning in the finance domain can change lives.

References: [1]

Problem Statement

Essentially, if you exclude more advanced investing and trading techniques, such as options or x, the main premise is to "buy low, sell high". That is, buy your investment/stock at a low price and hope to sell the same investment, some time in the future, at a price higher then what you bought it.

The problem to be solved: *Predict the future price of a given stock, given the historic prices of said stock, over a time period using concepts and techniques in technical analysis and machine learning.*

Datasets and Inputs

For this project, stock price indicator, the dataset to be used will be that of publicly traded companies from the Nasdaq stock market and NYSE (New York Stock Exchange) obtained for free from Yahoo! Finance via Python module "yahoo-finance".

Nasdaq and NYSE because of their size, number of companies listed on both exchanges and historic data available for training. In order to predict future stock prices, models would need current and historic data to to determine data behavior over time or similar characteristics.

Historic stock data will be used for EDA, feature selection and engineering, model training and finally the trained model will be used to predict future stock prices.

Dataset Characteristics

Below you can see a sample of features, values and definitions of which may be used as dataset and inputs for this project.

Name	Value	Description
Adj_Close	35.83	the closing price that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open
Close	35.83	the final price at which a security is traded on any given trading day
Date	2014-04-29	calendar date of the reported values
High	35.89	today's high is the highest price at which a stock traded during the course of the day
Low	34.12	today's low is the lowest price at which a stock traded during the course of the day
Open	34.37	the opening price is the price at which a security first trades upon the opening of an exchange on a given trading day
Symbol	YHOO	a stock symbol is a unique series of letters assigned to a security for trading purposes
Volume	28720000	the number of shares/contracts traded in a security during a given trading day
FullTimeEmployees	12200	Current number of full time employees
Sector	Technology	a sector is an area of the economy in which businesses share the same or a related product or service
start	1996-04-12	initial offering date at which a security is first made available for public purchase

References: [2]

Solution Statement

For this project, your task is to build a stock price predictor that takes daily trading data over a certain date range as input, and outputs projected estimates for given query dates.

A training interface that accepts a data range (start_date, end_date) and a list of ticker symbols (e.g. GOOG, AAPL), and builds a model of stock behavior. Your code should read the desired historical prices from the data source of your choice.

A query interface that accepts a list of dates and a list of ticker symbols, and outputs the predicted stock prices for each of those stocks on the given dates. Note that the query dates passed in must be after the training date range, and ticker symbols must be a subset of the ones trained on.

Potential solution - supervised learning model trained on historic data to predict, within a margin of error, the future closing price of a stock. This knowledge would put the owner of said information in an advantageous position, allowing him or her to invest in a company's stock increase or decrease in value.

Benchmark Model

The benchmark model will be comprised of multiple sources such as a naive random forest model and stock-forecasting.com an online stock price indicators. Outside of the train/test data backtesting will also be performed.

- Similar stock mark indicator model for benchmarking – http://www.stock-forecasting.com/Content/Data/Test.aspx
- A naive simple solution such as k-means or decision tree
- Backtesting – http://www.investopedia.com/terms/b/backtesting.asp – https://www.quantstart.com/articles/Backtesting-a-Forecasting-Strategy-for-the-SP500-in-Python-with-pandas

Evaluation Metrics

The benchmark model and solution model will be evaluated, when appropriate, with the Root Mean Squared Error and Rˆ2 score metrics.

Root Mean Squared Error (RMSE)

The root mean squared error or deviation measures the difference between values predicted by the model and the values actually observed. It is considered to be one of the most popular metrics for evaluating regression models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R Squared (Rˆ2)

R Squared, the coefficient of determination, is an indication of the goodness of fit of predicted values to the actual values.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Model results will be considered satisfactory if the predicted stock value 7 days in the future is within +/- 5% of the actual value, on average.

References: [3]

Project Design

The approach for this problem will, as a whole, take on the view of a product and have not just a predictive model but a user interface (UI) for someone to easily interactive with and request predictions. Additionally, the system will provide its own historical predictions and its final accuracy.

Below are the associated processes for both the 1) Modeling and 2) Product workflows:

Modeling Workflow - Supervised Learning

- Data gathering + collection
- Data pre-processing
- Feature engineering, missing data
- EDA and visualizations
- Model evaluation / cross-validation
- Feature reduction / selection
- Dimensionality reduction
- Train/test data split
- Tuning / hyperparamater optimization
- Final model
- Test dataset / backtesting
- Predictions
- New / future data

The modeling workflow will, for the most part, follow the process as put forward by Sebastian Raschka.

Data gathering and pre-processing

Data will be sourced mainly from Yahoo Finance API via the python module yahoo-finance 1.3.2. The data may be cleaned, formatted or missing values filled while new data, feature engineering may also be used such as the CAPM.

EDA and visualizations

In order to best fit and model the data, exploratory data analysis will first be performed to understand the type of data being used and determine any particular relationships or correlations. EDA and visualizations will take the shape of statistical data summary, log transformations, correlation matrix, etc. Once a handle of the data has been accomplished through analysis and graph plots the model evaluation phase will begin.

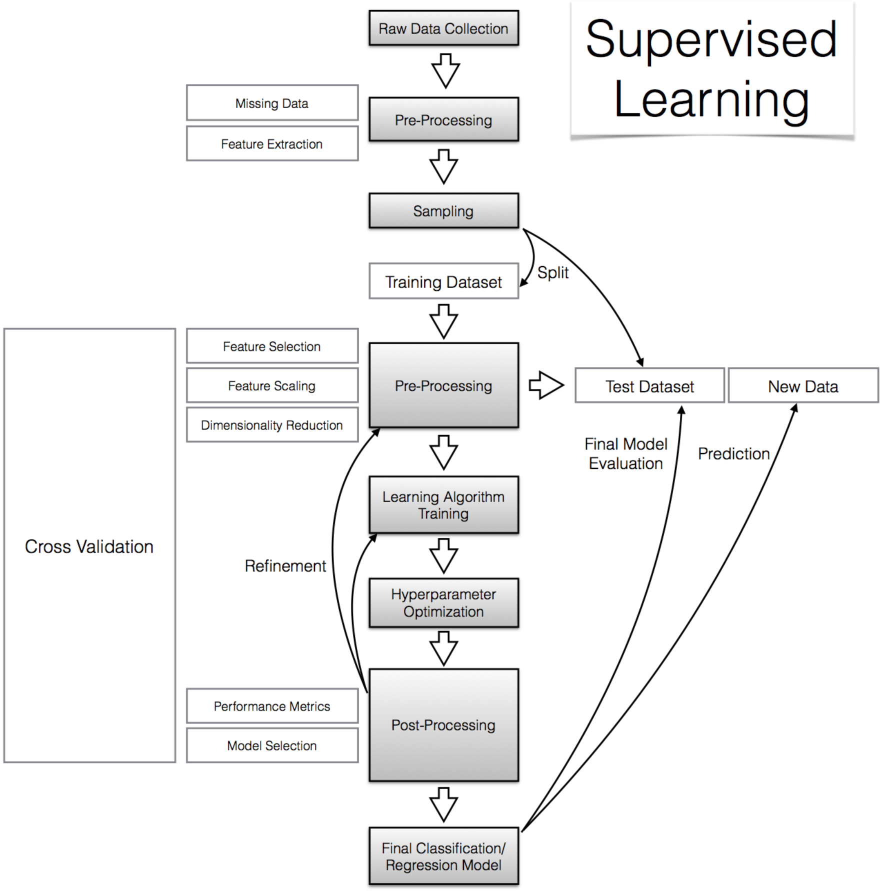
Raw Data Collection
Supervised Learning
Missing Data
Feature Extraction
Pre-Processing
Sampling
Split
Training Dataset
Cross Validation
Feature Selection
Feature Scaling
Dimensionality Reduction
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Refinement
Hyperparameter Optimization
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model

Figure 1:

Model evaluation, Cross-validation and Backtesting

Model evaluation will consider single regression as well as ensemble algorithms. While k-fold cross-validation will be used to evaluate the different combinations of feature selection, learning algorithms.

Evaluated metrics will be considered satisfactory if the predicted value is within +/- 5% of actual value.

Some of the algorithms do be considered are:

- Supervised Vector Machine (SVM)
- K Nearest Neighbour (KNN)
- Random Forest
- Adaboost
- Gradient Boosting

Product Workflow - Flask Microframework

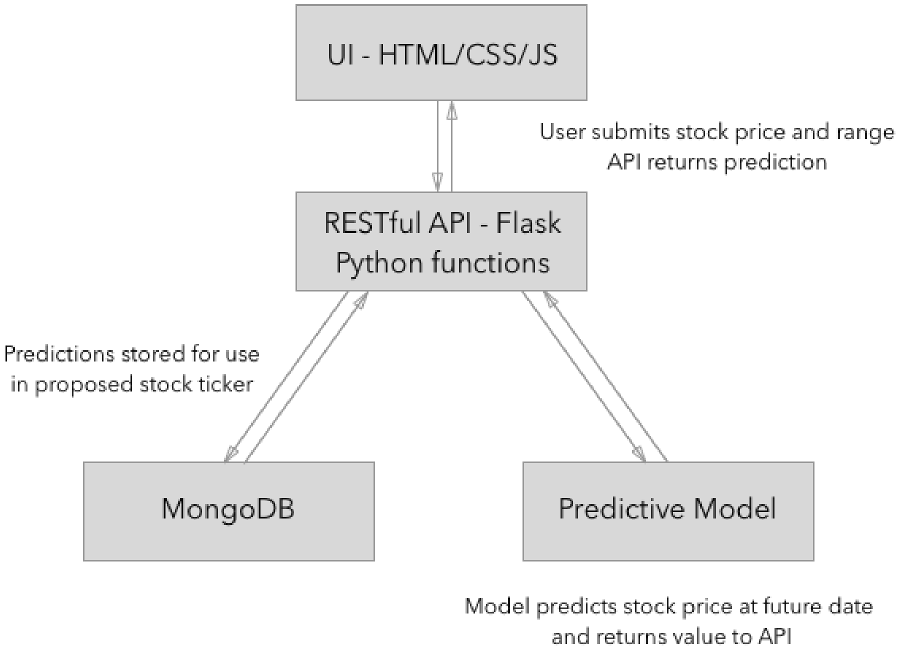

Figure 2:

Development perspective

- Environment setup
- Set up Flask
- Flask RESTful implementation

- Create restful endpoints
- Flask DB implementation
- Create data storage for tracking previous predictions and results
- Test APIs and data storage

User perspective

1. Visit landing page of stock price indictor
2. User enters stock symbol of stock price to predict
3. User selects a future range of prediction (e.g. 7 days from today)
4. The system will then display the stock price prediction

- The system will also display a "prediction stock ticker" which will display past predictions from previous user requests and the system's prediction along with actual results

References

1

- http://www.diva-portal.org/smash/get/diva2:354463/fulltext01.pdf
- http://www.academia.edu/11692137/Analyzing_Different_Machine_Learning_Techniques_for_Stock_M
- http://cs229.stanford.edu/proj2015/009_report.pdf
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.6139&rep=rep1&type=pdf
- http://cs229.stanford.edu/proj2012/ShenJiangZhang-StockMarketForecastingusingMachineLearningAlgori

2

- https://en.wikipedia.org/wiki/NASDAQ
- https://en.wikipedia.org/wiki/New_York_Stock_Exchange
- http://www.investopedia.com/terms

3

- https://en.wikipedia.org/wiki/Root-mean-square_deviation
- https://en.wikipedia.org/wiki/Coefficient_of_determination
- http://scikit-learn.org/stable/modules/model_evaluation.html
- https://www.kaggle.com/wiki/RootMeanSquaredError

4

- http://www.investopedia.com/terms/i/investment.asp
- http://bebusinessed.com/history/history-of-the-stock-market/
- http://www.investopedia.com/articles/07/stock-exchange-history.asp
- https://github.com/googledatalab/notebooks/blob/master/samples/TensorFlow%20Machine%20Learning
- https://github.com/rasbt/python-machine-learning-book

- http://francescopochetti.com/stock-market-prediction-part-introduction/
- https://www.quantstart.com/articles/Forecasting-Financial-Time-Series-Part-1
- http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html