Udacity Machine Learning Nanodegree

Project Proposal

Ashish Rao Mangalore

Figure 1[Source: https://techemergence.com/wp-content/uploads/2016/08/Machine-Learning-in-Finance.jpg]

Machine Learning in Finance

The project will be an approach to leverage the power of data to make predictions of a stock to the maximum possible accuracy. Investment funds are hedge funds are always on the lookout for processes to maximize profit. With advances in computational power and data processing techniques coupled with the ubiquitous nature of devices connected to the internet, machine learning based are extensively used to study the performance of stocks. There are generally two approaches to predicting the price of a stock in finance.

- Qualitative/Fundamental Analysis- Investors study terms sheets , profit and loss statements , company's credibility ,recent developments in the market and so on and then execute a trade. This approach to trading requires extensive reasoning and human intelligence and hence is not a good problem for machine learning to solve.

- The second approach is to use quantitative analysis to predict the future price of the stock based on various statistical parameters like rolling mean, daily returns, rolling standard deviation and so on. This approach makes use of data driven methods and would be a good problem to be investigated using machine learning.

Problem Statement

The objective of the project is to use machine learning algorithms to predict the behavior of a chosen stock. After observing the performance of multiple machine learning algorithms, the one giving the best results is selected. An applet is built in Python 2.7 for the same. Data pertaining to different company stock prices is available online ex. Quandl, Yahoo Finance, Google Finance and Bloomberg to name a few. These sources are processed using a suitable framework like Pandas and ML algorithms are built upon them.

This can be treated as a regression problem in supervised learning. The inputs to the ML based forecaster is going to be Time Series data , i.e date indexed integer data queried from online databases and the expected output is an integer which is the Adjusted Closing price of the selected stock for a forecast interval defined in the code.

Datasets and Inputs

The dataset for the purpose of this project will be the Stock of 'AAPL' or Apple. This is data is chosen because it has a lot of data points (right from the 1980's) and hence is a good problem for a machine learning based approach. Data is obtained through the Quandl API in Python 2.7. The data obtained is of the time-series type and it is processed using Pandas in Python. The Quandl API gives information about the opening price, closing price, daily high, daily low ,adjusted close and also the indication of the issue of dividends or stock splits (with a 1 or 0).The very basis of technical analysis is the prediction of stock prices by statistical approaches. However, the amount of insight a human can achieve is much less than what a machine can do given the same amount of data. Hence using a machine learning based predictor would be a very good way to go about with this dataset.

The features used for the machine learning algorithm can be the opening price, the All day high , the all day low , the 50 day rolling mean , the rolling standard deviation and so on between a selected date range .Further decision of the inputs are taken during the course of the project based on the performance of the model. The expected output is the Adjusted closing price of 'AAPL' for the queried date. The prediction here is to predict a continuous change in price and hence it is a regression problem. Other behaviors like direction of stock or bin wise classification are not predicted in this project.

Solution Statement

It is known that the objective of the project is to predict the adjusted close price of ‘AAPL’ given certain input features as required. Hence this mainly falls into the category of a regression problem. However, before a suitable regression algorithm can be applied, it is necessary that only those features be used that are relevant. This helps to reduce the training time and makes the model more robust. Hence feature selection is carried out by running trials with various combinations or by applying dimensionality reduction using PCA.

After the features have been appropriately reduced, a regression algorithm is applied to make predictions. Several algorithms like Static vector regression Nearest Neighbors, Neural Networks and decision tree regressors are available as methods in Scikit’s sklearn library.These methods function like black boxes and can be deployed by tweaking the parameters in them .After finding out the best functioning algorithm ,it is decided as a worthy model and benchmarked against other traditional methods of prediction.

Benchmark Model

Stock markets are known to be extremely stochastic and thus even with machine learning models the R2 score of the prediction is pretty low(in the range of 40-60%). If the models performed any higher than it would have been a case of an overfit and not a good sign of a predictor(otherwise you would be a billionare!!). Therefore getting a performance with predictor accuracies in the abovesaid range would amount to a good training for the stock in question. Another approach would be to fit a very rudimentary model like a linear regressor and compare the performance of the model with this very naïve regressor.

Evaluation Metrics

Being a regression project, the performance is evaluated by the score generated for the testing data for the queried date. These performance metrics are available as standards and need not be devised by the engineer from scratch. They indicate how well our algorithm is performing evaluation metrics commonly used is the R2 score.

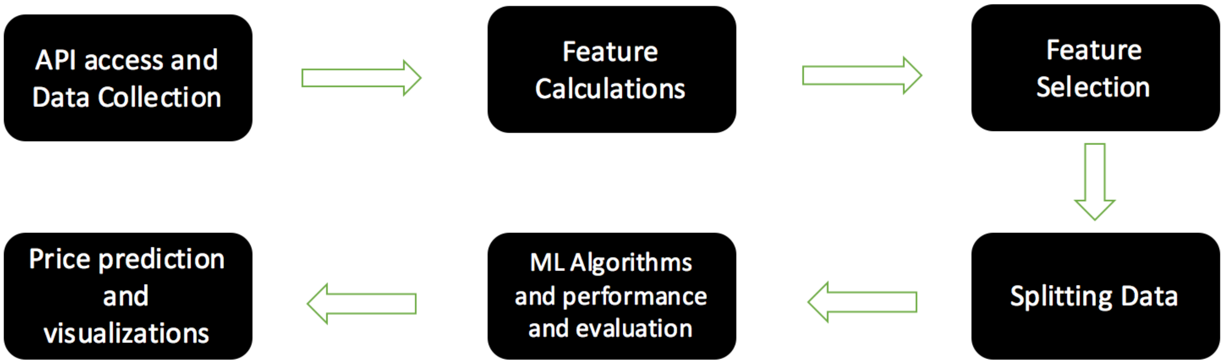
Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.

The coefficient R^2 is defined as (1 - u/v), where u is the regression sum of squares ((y_true - y_pred) ** 2).sum() and v is the residual sum of squares ((y_true - y_true.mean()) ** 2).sum().[2]

Based on the R2 score the engineer can compare different algorithms and see which one is more suited for their application. A higher R2 score will mean a better performing algorithm.

Project Design

The project is to implement a code to predict stock prices.It is implemented in python 2.7 in the Spyder IDE. It is divided into phases like accessing the API , pulling the data, dividing them to data frames, checking their authenticity and so on .It is explained in detail as follows


API access and Data Collection
Feature Calculations
Feature Selection
Splitting Data
ML Algorithms and performance and evaluation
Price prediction and visualizations

Stage 1: API access and data collection

Quandl is used to obtain data for the 'AAPL' stock. This is done by importing the Quandl API and setting up the configuration key. The data pertaining to a particular time period is called for and stored in a dataframe.The time period for training is specified and the data during this period is segregated in separate dataframes. Any anomalous data is looked for and removed from the data frame.

Stage 2: Calculating necessary features

While input features like 'Open', 'Close', 'High' , etc are readily available ,for the sake of improving the trainer , it becomes necessary to compute other features which can be incorporated into the training dataset. Features like 10 day rolling mean ,50 day rolling mean, daily returns are calculated and stored in data frames for training.

Stage 3: Feature Selection

The features which are correlated can be dropped as required as they add no value to the training. This can be done by a PCA or by selecting features manually and testing the performance of the model obtained.

Stage 4: Splitting the training and testing data set

The data is available in the form of time Series data and hence it is split by merely using the first N rows according to the training ratio specified (about 70%) or by using time series train test split methods.

Stage 5: Selection of Algorithm and Finalizing predictor

Since this is a regression problem various ML algorithms like KNN, linearfit , decision tree regressor , static vector regression and multi layer perceptron networks are run of the test and tested for .Additionally Cross validation can be run for better performance rather than just using one set of train and test data sets. The best performing algorithm is selected using performance metrics the R2 score.

Stage 6: Prediction of prices, Plotting Graphs and visualizations

The price of the stock is predicted for the queried date and all necessary graphs are plotted using matplotlib .It is possible to plot the performance of the algorithms used and the predicted prices in the queried date range. This helps in visual summarization of the prediction.