Stock Market Analysis

Ravi Ekambaram
Computer Science Department,
University at Albany, SUNY
ID: 00129953
rekambaram@albany.edu

Abstract— **Prediction of Stock Prices is not only inquisitiveness but also the very challenging topic. Developed countries' economies are measured according to their power economy. Currently, stock markets are an illustrious trading field because in many cases it gives easy profits with low risk rate of return. In this paper, we predict the stock market using k-nearest neighbor algorithm, to help executive, investors, user, and choice makers in making valuable decisions.**

Keywords— stock market, predictions, analysis, models, KNN, companies data

INTRODUCTION

Genereally there are 2 forms of data analysis to predicting future trends. Classification and prediction. Prediction model can be applied to any type of event, whether it occurred in past or is going to occur in future.

It is a process which uses probability and some other data mining techniques to forecast outcomes.

In prediction, each model is made up of a number of predictors, which are variables that are likely to influence future results. The procedure is general, like we collect the data first, Once data has been collected for relevant predictors(in this case companies), a model is formulated based on which we predict. The model can be a simple linear equation or it may be a complex network, mapped out by sophisticated software. As additional data becomes available, the analysis model is validated or revised.

MOTIVATION

Stock market is a kind of time series in financial domain, which change dynamically and selectively. Time series are difficult to predict, because the problem is nonlinear, non-stationary and have a lot of noises.

A type of prediction strategy, stationary strategy is also not possible, as investors will soon discover such strategies and those who are forecasting, their rules will lead to self-destruction. Thus, it will help the users to understand the stock prices and invest their hard-earned money accordingly to the stock market.

Potential users

This algorithm is helpful not only to the business man, but also to investors and general people who even invest a small amount of money in stocks.

Potential applications

This algorithm can be used to find nearest neighbor retrieval and classification, gene expression, text mining, agriculture, medicine, climate forecasting, Currency exchange rate, Bank bankruptcies, Understanding and managing financial risk, Trading futures, Credit rating, Loan management, Bank customer profiling, Money laundering analyses.

Problem statement

The rate at which the economy of developing countries is going low, it becomes necessary to invest the money in some safe companies so that it beneficial to users, companies, and the country. The Stock Market prediction task is interesting as well as divides researchers and academics into two groups those who believe that we can devise mechanisms to predict the market and those who believe that the market is efficient and whenever new information comes up the market absorbs it by correcting itself, thus there is no space for prediction.

Significance of Problem

As, we can see the problem is highly important, as the user will not know, what will be the future of the company that he is investing in.

Related work

There have been designed several different prediction systems for real world applications, using different models.

Some of the works are: Stock market prediction: A big data approach, predicting stock prices using data mining techniques, Application of data mining techniques in stock markets: A survey.

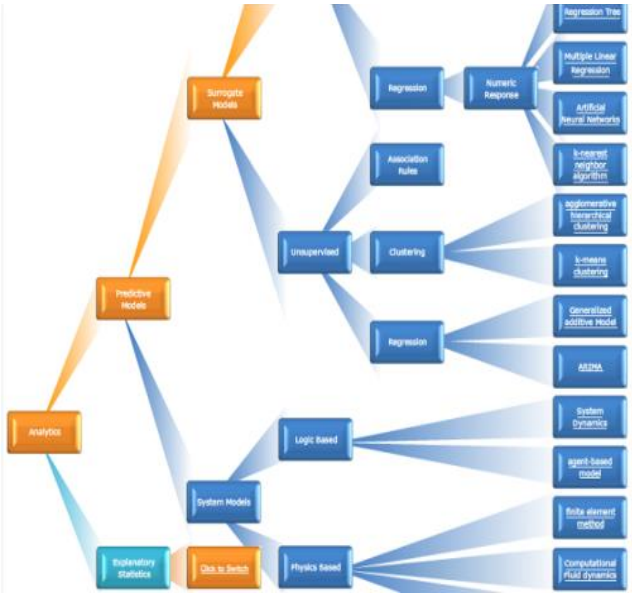
PROPOSED APPROACHES

Predictive models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set.

Depending on definitional boundaries, predictive modelling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts.

When deployed commercially, predictive modelling is often referred to as predictive analytics.

Different types of approaches used are:



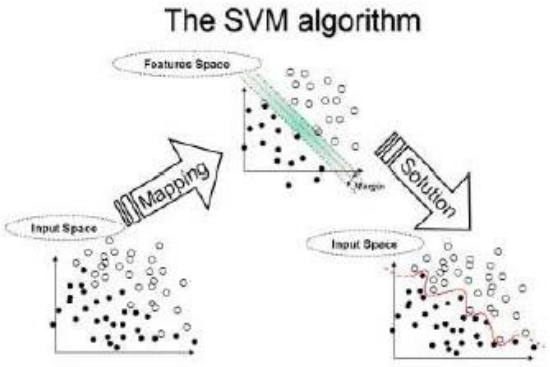
Usually, regression model is used for prediction purposes.

There are two classes of predictive models: parametric and non-parametric.

A third class, semi-parametric models, includes features of both.

Algorithms

- SVM: Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.



SVMs are generalized linear classifiers and can be interpreted as an extension of the perceptron. It has a special property, that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

Computing SVM classifier is of the form:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i + b)\right)\right] + \lambda\|w\|^2. \qquad (2)$$

We optimize it, using this formula:

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\zeta_i + \lambda\|w\|^2$$

We simplify it using:

$$\text{maximize } f(c_1 \ldots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i (x_i \cdot x_j) y_j c_j,$$

$$\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i.$$

- Second approach is decision tree: The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram in below. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

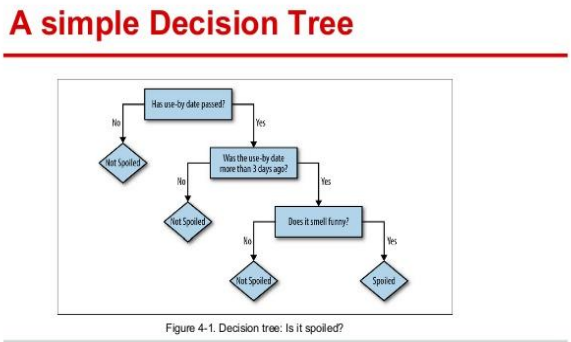


Figure 4-1. Decision tree: Is it spoiled?

Decision trees used in data mining are of two main types:

  - Classification tree analysis is when the predicted outcome is the class to which the data belongs.
  - Regression tree analysis is when the predicted outcome can be considered a real number. The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both above procedures.

- Logistic regression: It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).

  Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression estimates a multiple linear regression function defined as:

$$= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta$$

SYSTEM DESIGN AND IMPLEMENTATIONS

Data set: We collected our data set using RT stock API. It has values of stocks as per company. Their high values, low values, volume.

Size of data set is around 11 megabytes.

We have used the data as text file, thus, the size of data is not compressed either.

Data set:

{"Volume": "1464400", "Adj_Close": "177.523886", "High": "179.800003", "Low": "177.759995", "Date":
: "178.679993", "Open": "179.25"}{"Volume": "1804600", "Adj_Close": "177.335128", "High": "179.440002",
"Date": "2017-01-18", "Close": "178.490005", "Open": "177.899994"}{"Volume": "1557500", "Adj_Close":
"177.679993", "Low": "176.25", "Date": "2017-01-17", "Close": "177.259995", "Open": "177.00"}{"Volume":
": "176.242239", "High": "177.910004", "Low": "176.830002", "Date": "2017-01-13", "Close": "177.389999",
{"Volume": "1321800", "Adj_Close": "176.291919", "High": "177.699997", "Low": "175.75", "Date":
: "177.440002", "Open": "176.970001"}{"Volume": "1579600", "Adj_Close": "176.739004", "High":
"176.389999", "Date": "2017-01-11", "Close": "177.889999", "Open": "176.630005"}{"Volume": "2030100",
483", "High": "177.490005", "Low": "176.309998", "Date": "2017-01-10", "Close": "176.580002", "Open":
": "1617800", "Adj_Close": "176.123021", "High": "178.380005", "Low": "177.199997", "Date": "2017-01-09"
, "Open": "178.369995"}{"Volume": "1625000", "Adj_Close": "177.076801", "High": "178.600006", "Low":
"2017-01-06", "Close": "178.229996", "Open": "177.289993"}{"Volume": "1447800", "Adj_Close":
"179.139999", "Low": "176.889999", "Date": "2017-01-05", "Close": "177.710007", "Open":
": "1542000", "Adj_Close": "177.16623", "High": "178.899994", "Low": "177.610001", "Date": "2017-01-04",
, "Open": "178.029999"}{"Volume": "2509300", "Adj_Close": "176.897973", "High": "180.00", "Low":
"2017-01-03", "Close": "178.050003", "Open": "178.830002"}{"Volume": "1594200", "Adj_Close":
"179.479996", "Low": "178.289993", "Date": "2016-12-30", "Close": "178.570007", "Open":
1102000", "Adj_Close": "177.255644", "High": "179.139999", "Low": "178.029999", "Date": "2016-12-29",
, "Open": "178.289993"}{"Volume": "1287900", "Adj_Close": "176.927777", "High": "179.449997", "Low":
"2016-12-28", "Close": "178.080002", "Open": "178.880005"}{"Volume": "651000", "Adj_Close":
"179.199997", "Low": "178.570007", "Date": "2016-12-27", "Close": "178.919998", "Open":
": "731100", "Adj_Close": "177.59344", "High": "179.289993", "Low": "178.529999", "Date": "2016-12-23",

Graphic user interface

We do not any GUI. We read the input file in form of a text file and using KNN model we tried to get the output in the form of excel spread sheet, predicting the future values.

Major components:

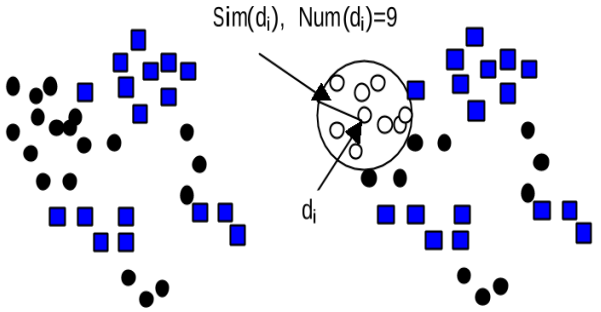
The main components in this system are stock prices. High and low values of stocks which have been gathered. The algorithms and implementation uses KNN model to give us the future values of the input stocks.

MODEL

Item-based K-nearest neighbor (KNN) algorithm: Its philosophy is as follows: The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

$$sim(a,b) = \frac{\sum_{u \in U(a) \bigcap U(b)} (R_{a,u} - \bar{R}_u)(R_{b,u} - \bar{R}_u)}{\sqrt{\sum_{u \in U(a) \bigcap U(b)} (R_{a,u} - \bar{R}_u)^2 \sum_{u \in U(a) \bigcap U(b)} (R_{b,u} - \bar{R}_u)}}$$



The advantage of the above-defined adjusted cosine similarity over standard similarity is that the differences in the rating scale between different users are taken into consideration.

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification. A good *k* can be selected by various heuristic techniques.

The special case where the class is predicted to be the class of the closest training sample (i.e. when *k* = 1) is called the nearest neighbor algorithm.

As its name indicates, KNN finds the nearest K neighbors of each movie under the above defined similarity function, and use the weighted means to predict the rating.

For example, the KNN algorithm for movies leads to the following formula:

$$P_{m,u} = \frac{\sum_{j \in N_u^K(m)} sim(m,j) R_{j,u}}{\sum_{j \in N_u^K(m)} |sim(m,j)|},$$

How does KNN model works in our algorithm?

Future stock market closing price is computed by:

i. Determine the number of nearest neighbors, k.

ii. Compute the distance between the training samples and the query record.

iii. Sort all training records according to the distance values.

iv. Use a majority vote for the class labels of k nearest neighbors, and assign it as a prediction value of the query record

- To formulate the EM algorithm formula in this case, let latent random variable Gm ~ Qm(·) denotes the group of stocks s. Thus, the E-step formula is described as follows:

$$Q_m^{(t+1)}(g) = P(G_m^{(t+1)} = g | R_{u,m}; \mu_{g,u}, \sigma_{g,u}^2) = \frac{Q_m^{(t)}(g) \Pi_{u \in U(m)} \frac{1}{\sqrt{2\pi}\sigma_{g,u}} exp\left(-\frac{(R_{u,m}-\mu_{g,u})^2}{2\sigma_{g,u}^2}\right)}{\sum_{g'} Q_m^{(t)}(g') \Pi_{u \in U(m)} \frac{1}{\sqrt{2\pi}\sigma_{g',u}} exp\left(-\frac{(R_{u,m}-\mu_{g',u})^2}{2\sigma_{g',u}^2}\right)},$$

$$\max_{\mu_{g,u},\sigma_{g,u}^2} \sum_m \sum_g Q_m(g) \left[log\left\{ \Pi_{u \in U(m)} \frac{1}{\sqrt{2\pi}\sigma_{g,u}} exp\left(-\frac{(R_{u,m}-\mu_{g,u})^2}{2\sigma_{g,u}^2}\right)\right\} - log\left(Q_m(g)\right)\right],$$

$$\max_{\mu_{g,u},\sigma_{g,u}^2} \sum_{m,g} Q_m(g) \sum_{u \in U(m)} \left[-log(\sigma_{g,u}) - \frac{(R_{u,m}-\mu_{g,u})^2}{2\sigma_{g,u}^2} \right].$$

$$\mu_{g,u} = \frac{\sum_{m \in M(u)} Q_m(g) R_{u,m}}{\sum_{m \in M(u)} Q_m(g)}$$

$$\sigma_{g,u}^2 = \frac{\sum_{m \in M(u)} Q_m(g) (R_{u,m}-\mu_{g,u})^2}{\sum_{m \in M(u)} Q_m(g)}$$

$$P_{u,m} = \sum_g Q_m(g) \mu_{g,u}.$$

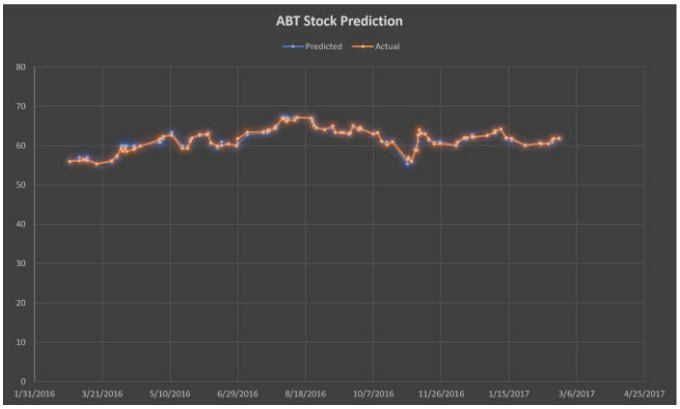
OUTPUT

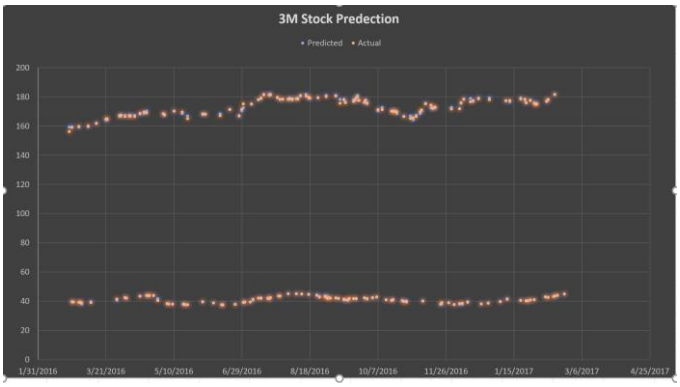
This is how the spreadsheet looks like:

Date	Predicted	Actual
########	61.66	61.83
########	61.66	61.77
########	60.8	61.48
########	60.42	60.51
2/8/2017	60.42	60.52
2/7/2017	60.67	60.56
########	60.17	60
########	61.27	61.86
########	61.66	61.99
1/9/2017	64.1	64.21
1/5/2017	63.27	63.77
1/4/2017	63.39	63.29
########	62.5	62.62
########	62.22	62.16
########	62.71	62.22
########	61.66	62.02
########	61.67	61.99
########	60.7	60.9
########	60.18	59.99
########	61	60.51

Date	Predicted	Actual
########	120.94	122.1
########	120.94	121.23
########	120.76	120.72
########	119.68	120
2/8/2017	115.5	115.43
2/6/2017	115.04	114.18
2/3/2017	113.95	114.49
2/1/2017	114.6	113.21
########	113.6	113.87
########	115.44	115.35
########	116.5	115.93
########	116.35	116.48
########	115.35	114.82
########	115.5	115.07
1/6/2017	116.35	116.3
1/4/2017	117.23	116.74
########	117.45	117.55
########	117.45	117.48
########	124.1	123.65
########	124.1	123

Comparison





The above figures are the comparisons between the predicted values of stock using our model and the real values for three companies: ABBV, 3M, and ABT.

CONCLUSION

We have found the output, which is near to the values of the stock, which the companies had at the end of the day.

But the limitation is that because we have used "*RTstock API*" we weren't able to get enough data.

The above API does not give any information about the external factors affecting the prices, that's going to be a downfall for our prediction but we are not expecting a major one.

As the current prediction is based on historic stock prices, We have planned to retrieve some news about that company on that given date and analyze the sentiment and try predicting the stock of that company.

REFERENCES

1) http://acit2k.org/ACIT/2013Proceedings/163.pdf
2) Stock market prediction: A big data approach
3) http://www.jobmonkey.com/daytrading/stock-analysis/
4) https://en.wikipedia.org/wiki/Predictive_modelling
5) https://www.rroij.com/open-access/prediction-using-back-propagation-and-knearest-neighbor-knn-algorithm.pdf
6) https://www.diva-portal.org/smash/get/diva2:771141/FULLTEXT01.pdf
7) https://www.researchgate.net/publication/228664309_Application_of_data_mining_techniques_in_stock_markets_A_survey
8) http://searchdatamanagement.techtarget.com/definition/predictive-modeling
9) https://www.researchgate.net/publication/224385407_A_Data_mining_algorithm_to_analyse_stock_market_data_using_lagged_correlation
10) Applying Data Mining Techniques to Stock Market Analysis
11) Gabriel Fiol-Roig, Margaret Miró-Julià, and Andreu Pere Isern-Deyà*