Study Project Report

on

Predicting Stock Prices using Machine Learning

Submitted by

Angad Lamba(2014B3A7689P)
Megha Shishodia(2014B3A7605P)

To

Prof. N. V. M Rao
Department of Economics & Finance
BITS Pilani, Pilani Campus

Table of Contents

S. NO.	CONTENT	PAGE NO.
1	Abstract	2
2	Background and Scope	3
3	Implementation	4 - 9
4	Conclusion	10
5	References	11

I. ABSTRACT

Financial market is a very intricate and dynamic system. The market is constantly changing and is influenced by day to day economic, political and social activities. Hence it is a very intriguing field of research for investors and research scholars. It seems that the market is unpredictable and random but clear patterns can be extracted using various mathematical models and machine learning algorithms. But estimating parameters in the financial market can be a cumbersome task, keeping in mind the large number of factors influencing them. Our project focuses on predicting these parameters and their direction of movement using some of the determinants of stock market. Machine learning techniques for classification namely Support Vector Machine and Naïve Bayes classifier along with other classification methods, have been used. The model is trained on old stock market data and has been found to give results with significant accuracy, when tested on new data. The results of the model are instrumental in designing profitable trading strategies in the stock market. In this project we attempt to predict whether a stock price sometime in the future will be higher or lower than it is on a given day. We find little predictive ability in the short-run but definite predictive ability in the long-run. The accuracy derived for each regression model tells us how well the model predicts the stock price.

II. BACKGROUND AND SCOPE

Stock market is one of the most dynamic and variable field, which fluctuates with economic changes and political decisions. This intrinsic characteristic of stock market interests many investors, yearning to make huge profits and researchers, looking for scientific patterns. Many data scientists have tried to apply various machine learning models, trained on stock data and observed the results produced. Though the results produced may not be convincing but are somewhat in accordance with various economic theories. Our paper aims to do the same with hope of giving better results.

III. IMPLEMENTATION

1. Data Collection:

The training data used in our project were collected from free database available at quandl.com. In this project, we picked BSE-SENSEX to apply our methods. The data contains daily stock information ranging from 03/04/1979 to 24/03/2017 (8702 data points).

- The data set consist of following features initially: Open, High, Low, Close, where they mean the following:
  Open: A particular day's opening price
  Close: A particular day's closing price
  High: A particular day's highest price
  Low: A particular day's lowest price
- These features are used in our analysis to extract new features which are:
  Open1: A particular day's previous day's opening price
  Close1: A particular day's previous day's closing price
  High1: A particular day's previous day's highest price
  Low1: A particular day's previous day's lowest price
  Change: A particular day's closing and opening price difference
  Change1: A particular day's previous day's closing and opening price difference
- Features that were selected in our final model were:
  Open, High1, Low1, Change1, Close1.

In addition, we used the daily labeling as follows: label "1" if the closing price is higher than that of the opening price of that day. Otherwise label "-1" i.e. if Change > 0, label is "1" else "-1".
For example: if closing price of stock A on a particular day is 1000 and opening price is 900 then Change is 100 i.e. label is "1".

Stock	BSE-SENSEX
Data Source	Quandl.com
Features	Open, High1, Low1, Change1, Close1.

2. Model Selection:

In our project, we mainly applied supervised learning theories, i.e. Logistic Regression, Naive Bayes Classifier, Decision Tree, and SVM. The most important result that we should watch closely is the accuracy of prediction, which we define as follows:

$$Accuracy = \frac{the\ number\ of\ days\ that\ model\ correctly\ classified\ the\ testing\ data}{total\ number\ of\ testing\ days}$$

We used 80% of the data sets as training data and tested our fitted models with the remaining 20% data sets.

i. 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. It is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

ii. Decision Tree Learning: Decision tree learning uses a decision tree as a predictive model observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

iii. Naive Bayes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

iv. Logistic Regression: Logistic Regression is a regression model where the dependent variable (DV) is categorical. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

We noticed that there were some missing values, in our data, which posed a challenge. So we tried out different methods to fill out the missing data and the four models described above were used.The results were compared across different models and methods.

A. Fill by Zero:

The missing values were replaced by filling zero in those places. The following results were obtained:

Model	Logistic Regression	Naive Bayes Classifier	SVM	Decision Tree
Accuracy	64.84%	65.13%	60.94%	64.56%

It turns out that Naive Bayes Classifier predicts with the greatest accuracy. Followed by Logistic Regression, if the missing values are filled by zeroes.

B. Mean:

The missing values were replaced by filling mean value of the available values. The following results were obtained:

Model	Logistic Regression	Naive Bayes Classifier	SVM	Decision Tree
Accuracy	65.93%	60.65%	66.39%	63.98%

In this case, when the missing values are replaced by the mean, Support Vector Machine gave the best results with an accuracy of 66.39%. SVM was closely followed by Logistic Regression.

C. Median:

The missing values were replaced by filling median value of all the available values. The following results were obtained:

Model	Logistic Regression	Naive Bayes Classifier	SVM	Decision Tree
Accuracy	65.88%	60.59%	66.39%	64.56%

When the missing values were replaced by the median, Support Vector Machine, once again gave the best results. Followed again, by Logistic Regression.

D. Mode:

The missing values were replaced by filling mode value of all the available values. The following results were obtained:

Model	Logistic Regression	Naive Bayes Classifier	SVM	Decision Tree
Accuracy	65.76%	60.59%	66.39%	64.44%

Support Vector Machine proved to be the best for this case as well, And Logistic Regression once again produced the second best accuracy.

3. Results:

The following results can be observed when the final models were trained on our stock dataset, with the selected features used in our analysis:

1. The SVM gives the best result in terms of test accuracy (when missing values were filled by method B, C, & D.), of about 66.39%, as compared to other models used.
2. Naive Bayes on average performs worst among our selected models.
3. Among our 4 methods(A, B, C & D) of filling missing values in our data, Method-C i.e. median method, gives best results on an average, for our selected models.
4. Method-A i.e. fill by zero method, gives the most variable result in terms of test accuracy.

IV. CONCLUSION

In this project, we applied supervised learning techniques in predicting the stock price trend of a single stock. Support Vector Machine proved to give the best prediction of the stock market. This result is concurrent with the analysis carried out by most of the research work done in this field. On comparing the results obtained by using different methods to fill the missing values, we concluded that median should be preferred over other methods, to obtain the best results. Also, fill by zero method, gives the most variable result in terms of test accuracy and thus should be discarded in future analysis. When we carried out the analysis on a smaller dataset the accuracy was very poor(less than 50%). Whereas larger datasets gave a much better accuracy. This proves that as more information is available to the buyer, there is more probability of him predicting the market movement correctly. Hence semi market hypothesis doesn’t hold for limited information availability, but when the information is abundantly available to the buyer then the hypothesis holds true.

V. References

[1] "S&p dow jones indices," PDF, March 31 2015.

[2] C. Y. Z. Ben Jacobsen, "Are monthly seasonals real? a three century perspective," 2012.

[3] K. jae Kim, "Financial time series forecasting using support vector machines," Neurocomputing, vol. 55, 2003.

[4] G. F. Bjoern Krollner, Bruce Vanstone, "Financial time series forecasting with machine learning techniques: A survey," in European Symposium on Artificial Neural Networks: Computational and Machine Learning, Bruges, Belgium, April 2010.

[5] J. Brownlee. A tour of machine learning algorithms. [6] M. Buhmann, "Radial basis function." Mathematisches Institut, Justus-Liebig-Universität Giessen, 2010