International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Survey of Supervised Learning Techniques On Stock Trend Prediction

Ganesh Bhat[1], Aman Mathur[1], Karthik [1], Akash P H[1],Kavya N L[2]

[1]Student, Dept. of Computer Science Engineering, Sapthagiri College Of Engineering, Karnataka, India
[2]Professor, Dept. of Computer Science Engineering, Sapthagiri College Of Engineering, Karnataka, India

---

Abstract - *Stock Market the place where one can trade stocks, equities, bonds, etc. Trading in the stocks is profitable but the value of the stock is prone to change but if done right it is a highly profitable. Hence being able to predict the trend of the market allows gives a person to gain an edge in the market. This paper evaluates how the supervised learning methods such as Support Vector Machines, Random Forest and Multinomial Naïve Bayes perform when trying to accurately predict the trend of the market. For these models we process the raw data and generate a feature set. We also use various feature selection techniques on the feature sets to optimize the prediction models for comparison. By using feature selection we remove the noisy data to improve both accuracy and training time. Standard metrics such as accuracy, precision, and F1 score are used to draw a comparison between the prediction models created by these learning techniques.*

Key Words: Stock Market, Trend Prediction, SVM, Random Forest, Multinomial Naïve Bayes, Comparison

1. INTRODUCTION

According to Investopedia, the stock market refers to the collection of the of markets and exchanges where issuing and trading of the equities or stocks of publicly held companies, bonds and other classes of security take place. Stock market is a volatile and complicated system. Stock trend forecasting is one of the most challenging tasks which gives an edge to the investor while trading in the stock market. To effectively get accurate trends and reduce the risks we use machine learning models. This paper proposes the use of supervised machine learning models such as that of Support Vector Machine, Random forest and Multinomial Naïve Bayes algorithms to develop a trend predictor program.

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers' detection. It approximates a non-linear separating boundary by multi-local linear classifies with interpolation. This helps prevent the over fitting of the trend of the stock market. Random Forest is a flexible, easy to use machine learning algorithm provide good classification results even without hyper-parameter tuning. This is because in random forest classifier we can generate a large number of trees, the higher the number of trees in the forest gives the high accurate results. Multinomial Naïve Bayes is based on Bayes' probability theorem. It widely used in text classification which contain high dimensional training data sets. It is not only known for its simplicity, but also for its effectiveness. Using Naive Bayes algorithm one can quickly train and build prediction models.

The parameters like open, close, volume, high, low and adj close are scraped from the web for a given stock. Yahoo finance is one of the popular websites to get the required data for a given stock. We also consider two market indexes as well. We then prepare the other financial indicators such as relative strength index, rate of change, simple moving average that can be used to construct the dataset.

Once the dataset has been prepared, feature selection is performed to create the feature set. The prediction models are constructed using the feature sets along with hyper parameter optimization. These prediction models can get the stock the trend of the given market.

2. METHODOLOGY

This section gives a brief overview of some of the popular machine learning algorithms such as Support Vector Machine, Random Forest Classifier, and Multinomial Naïve Bayes. This paper studies the performance of these algorithms for stock market trend prediction.

2.1 Support Vector Machine

Support vector machines (SVMs) are supervised learning methods which are used for classification, regression and outlier detection. A Support Vector Machine classifies the data by creating hyper plane. In other words, for a given training data the algorithm creates an optimized hyper plane which is used to categorize new examples.

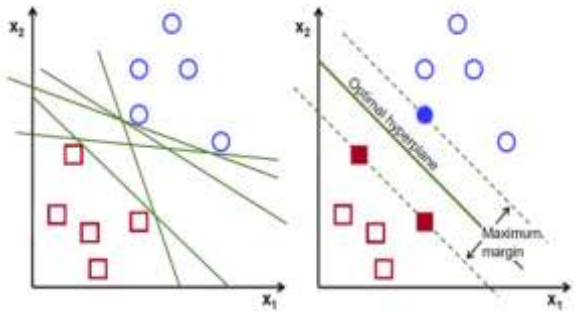

Fig. 1: [1]Possible hyper planes(left), Optimal hyper plane(right).

There are many possible hyper planes that could be chosen to separate the two classes of data points. A plane has to be found that has the maximum margin, i.e., the maximum distance between data points of both classes as shown in Fig-1. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

A hyper plane is a subspace of one dimension less than its ambient space .If w is a vector and b is a bias the equation is given by:

$$\mathbf{w^t \cdot b = 0}$$

Due to the reasons given in [5], C=14.02 and gamma=0.11 are considered as the hyper parameters while creating the SVM prediction model.

2.2 Random Forest Classifier

Random Forest is a supervised learning algorithm. As the name suggests, a forest is created which is made random. The forest built is an ensemble of Decision Trees, most of the time trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.Fig-2 shows the depiction of the random forest classifier.

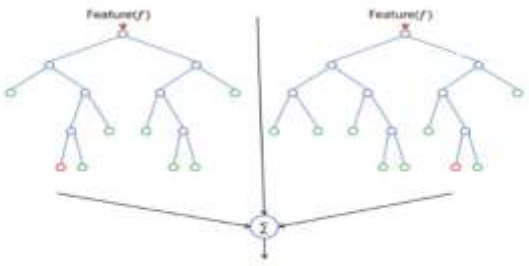

Fig. 2: [2] A random forest with two trees

Random Forest also adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. It is to be noted that in random forest classifiers the final result is obtained after performing a majority voting on the results obtained from the sub-trees.

2.3 Naïve Bayes

Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier. The Naive Bayes algorithm is called naive because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features.

It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. In spite of the naïve design and oversimplified assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems.

The basis of Naive Bayes algorithm is Bayes' theorem or alternatively known as Bayes' rule or Bayes' law. It gives us a method to calculate the conditional probability, i.e., the probability of an event based on previous knowledge available on the events. More formally, Bayes' Theorem is stated as the following equation:

$$\mathbf{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

Another popular variation of Naïve Bayes is Multinomial Naïve Bayes algorithm. This algorithm estimates the conditional probability of a particular word/term/token given a class as the relative frequency of term in documents belonging to class and is given by the equation:

$$\mathbf{P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}}$$

Thus this variation takes into account the number of occurrences of term t in training documents from class c, including multiple occurrences. Usually Multinomial Naive Bayes is used when the multiple occurrences of the words matter a lot in the classification problem. Such an example is when we try to perform Topic Classification.

3. Experiment and Evaluation

3.1 Dataset

The data for the stocks is obtained from Yahoo Finance. The data can also be mined from any other source as well. The dataset contains raw data such as Open Prices, Volume, Close Prices, Adj Close Price, High Price and Low Price. The data obtained is then pre-processed to generate

different stock market indicators which can be considered as features such as Moving Average, Rate Of Change, etc. This is because these indicators give a better picture of such as how the prices are shifting and what is the rate at which the price is changing. The indicators considered in the dataset are given in Table 1. These indicators are very popular and have extensively been used for trend analysis. We also consider a stock market index such as NASDAQ, S&P500, etc. like in [2]. The relevant stock market index is selected for a given stock. This gives us enough data for training our models as well. For the dataset we have pulled the data from 4/28/2016 to 4/28/2019.

Indicators
Open Prices
Close Prices
High Prices
Low Prices
Adjusted Close
Volume
Williams %R
Rate Of Change
Market Momentum
On Balance Volume
Relative Strength Index
Simple Moving Average
Commodity Channel Index
Average True Range
Money Flow Index
Exponential Moving Average

Table 1: Indicators considered as features for the dataset.

3.2 Experimental Setup

Figure 3 provides a general overview of how a trained model is created for this experiment.

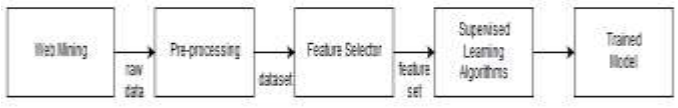

Fig. 3: General overview of creating a prediction model.

Before analyzing the algorithms, we must first create the labels for dataset. These labels determine whether the trend is moving upwards or downwards. For the experiment presented in the paper the trend was determined for the very next day. In this experiment, the trend is considered to move upwards if the present day open price is less than the open price of the next day.

We use web mining techniques to obtain the raw data for the stocks of a company. Here we use Yahoo Finance to obtain the required data. Next, pre-processing of the data is done. This involves obtaining other stock indicators to create the dataset. The dataset is then scaled to make sure all the features are in a standardized range. Feature selection is performed on the dataset using various feature selection techniques such as SelectBestK and ExtraTreeClassifier to obtain feature sets. These feature sets are split into training and testing data where 80% is used for training and 20% is used for testing. These feature sets are used to train each of the supervised learning models and tested using the testing data. The results are then tabulated and presented in the result section.

3.3 Evaluation Metrics

To evaluate the performance of the supervised learning methods considered in the paper. We use metrics such as Accuracy, Precision, Recall and F1-Score.

Accuracy: It is the fraction of instances that are classified correctly over the total amount of relevant instances.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: It is the fraction of relevant instances among the retrieved instances.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

Recall: It is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

Where, TP-True Positive FP- False Positive
TN-True Negative FN-False Negative

F1-Score: It can be defined as a measure to test the accuracy of a given test. It is computed using both precision and recall.

$$\textbf{F1-Score} = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

3.4 Results

The tabulated results are in percentage.

Company Name	Support Vector Classifier	Random Forest Classifier	Multinomial Naïve Bayes
Acitivison	83.67	81.63	78.23

Microsoft	70.74	72.78	68.02
Amazon	76.87	73.46	78.23
Apple	79.59	84.35	73.46
Netflix	81.63	79.59	75.51

Table 2: Accuracy Of Supervised Learning Techniques when no feature selection is applied

Company Name	Support Vector Classifier	Random Forest Classifier	Multinomial Naïve Bayes
Acitivison	83.67	81.63	80.95
Microsoft	71.42	72.78	68.02
Amazon	78.91	73.46	78.23
Apple	79.59	84.35	73.46
Netflix	81.63	79.59	80.27

Table 3: Accuracy Of Supervised Learning Techniques when Random Forest selection is applied

Company Name	Support Vector Classifier	Random Forest Classifier	Multinomial Naïve Bayes
Acitivison	83.67	81.63	81.632
Microsoft	73.46	78.23	70.06
Amazon	78.91	80.27	79.59
Apple	83.67	83.67	76.69
Netflix	81.63	82.31	81.63

Table 4: Accuracy Of Supervised Learning Techniques when SelectBestK selection is applied

Company Name	Support Vector Classifier	Random Forest Classifier	Multinomial Naïve Bayes
Acitivison	83.67	84.35	81.63
Microsoft	72.10	78.91	69.38
Amazon	78.91	78.90	80.95
Apple	83.67	85.03	76.19
Netflix	82.45	82.31	82.19

Table 5: Accuracy Of Supervised Learning Techniques when ExtraTreeClassifier selection is applied

Table 2 gives us the accuracy of the prediction model when no feature selection is applied. Tables 3, 4, 5 gives us the accuracy of the prediction model when feature selection is done. From the data we notice that Support Vector Machines and Random Forest Classifier are more robust to noisy data and gives us we good accuracy even when feature selection is not done while Multinomial Naïve Bayes is found to be very sensitive to noisy data as we can see significant jump in the accuracy.

Company Name	Support Vector Machine	Random Forest	Multinomial Naïve Bayes
Activison	**84.35**	81.63	80.95
Microsoft	73.46	**78.91**	70.06
Apple	83.67	**85.03**	76.19
Amazon	80.27	78.91	**80.95**
Netflix	**82.31**	81.31	81.63

Table 6: Accuracy of the Supervised learning Metrics

Company Name	Support Vector Machine	Random Forest	Multinomial Naïve Bayes
Activison	**97.91**	97.09	84.78
Microsoft	72.41	**77.01**	66.34
Apple	**84.33**	82.02	73.0
Amazon	**80.89**	79.56	78.12
Netflix	**87.01**	84.33	84.33

Table 7: Precision of the Supervised Learning Techniques

Company Name	Support Vector Machine	Random Forest	Multinomial Naïve Bayes

Activison	88.04	**89.13**	86.95
Microsoft	83.33	85.89	**88.46**
Apple	90.12	**91.35**	90.12
Amazon	85.71	88.09	**91.66**
Netflix	83.13	**84.33**	80.72

Table 8: Recall of the Supervised Learning Techniques

Company Name	Support Vector Machine	Random Forest	Multinomial Naïve Bayes
Activison	**87.43**	85.86	85.56
Microsoft	76.64	**81.21**	75.40
Apple	85.36	**87.05**	80.66
Amazon	83.61	83.23	**84.61**
Netflix	**84.95**	84.3	84.14

Table 9: F1-Score of the Supervised Learning Techniques

Tables 6, 7, 8 and 9 show the accuracy, precision, recall and F1-scores of the supervised learning techniques we have considered for the stock trend prediction of the next day. We see from the tables that SVM and Random Forest outperform Multinomial Naïve Bayes and the results between SVM and Random Forest are comparable. It can also be found from the results that SVM and Random Forest Classifier are more consistent than Multinomial Naïve Bayes.

4. CONCLUSION

This paper aimed at surveying the different supervised learning techniques for stock trend prediction. A comparison study has been performed between the prediction models created using SVM, Random Forest and Multinomial Naïve Bayes. A survey of the different feature selection techniques is also carried out and a thorough analysis is also performed on how they affect the performance. It was found that SVM and Random Forest outperforms Multinomial Naïve Bayes and they are less sensitive to noisy data as well.

REFERENCES

[1] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[2] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[3] https://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier

[4] https://www.hackerearth.com/blog/machine-learning/introduction-naive-bayes-algorithm-codes-python-r/

[5] Gangamol Jaiwang, Piyasak Jeatrakul," A Forecast Model for Stock Trading using Support Vector Machine", 2016 IEEE.

[6] Shashank Tiwari , Akshay Bharadwaj , Dr. Sudha Gupta ,"Stock Price Prediction Using Data Analytics" , IEEE 2017