Stock Market Analysis and Portfolio Optimization Using Ensemble Approach

Ayush Bhardwaj¹, Akshay Mahajan², Pranava Sureka³, N.Arivazhagan
Student, Student, Student, Assistant Professor
Information Technology, SRM Institute of Science and Technology, Chennai, India

Abstract: In this paper, we have built a predictive system for stock market by implementing machine learning based trading strategies like linear regression, KNN and regression trees including the algorithmic steps from information gathering to market orders. The focus is on applying probabilistic machine learning approaches to trading decisions. We use statistical approaches and apply them to trading scenarios. We analyze the historical data and clean it using statistical algorithms. We then take this clean data and train it with machine learning algorithms hence getting a prediction on the price of stocks. The developed tool can be used for in-depth analysis of stock market as well as can provide suggestions on managing one’s portfolio.

Index Terms: Linear Regression; Random-Forest; Ensemble; K nearest Neighbor; Sharpe-Ratio; Correlation.

I. INTRODUCTION

Stock Market Prediction has gained increasing importance in recent years as many corporate and non-corporate people, are interested to invest in this to gain profit. Many investors who invest in stock markets usually are not aware of the stock market behavior. If any system is developed which can consistently predict the trends of the dynamic stock market, would help the investor to gain more profit. Building accurate model is difficult because it possesses many theoretical and experimental challenges. Stock market prices depend on multiple factors such as social media data, fundamentals, and government bonds production of the company, historical prices, stock splits, news, and country’s economy. There have been many models which can predict the stocks prices. Building a model by considering only one factor might not be accurate. Hence integrating multiple factors historical prices and news might increase the accuracy of the model.

The goal of this research work is to build a model which include multiple modules to provide better results. The model uses supervised machine learning algorithms Random forest, Linear SVM, KNN which helps to determine when to buy a stock and when to sell it. The project will make attempt to deciding how much money to allocate to each prediction. This project will focus exclusively on predicting the daily trend (price movement) of individual stocks. The first module contains collecting the historical data of the company and manipulates it using statistical algorithms. The second module we have applied algorithms, methods, functions to manipulate and work with financial data. In the last module, we work with the data which we analyzed in previous modules and use it with machine learning algorithms to build a trading algorithm.

II. BACKGROUND AND LITERATURE SURVEY

In the recent developments, we have seen that the shortest path algorithm has found its popularity in the field of research but the safest path algorithm is as necessary as the shortest path algorithm. With the interest of researchers in the deployment of the safest path algorithm, we found that our main objective is to provide the trade-off between safety and distance. Moving to the background of the safest path algorithm, we have noticed that the crime zone areas are the only attribute which is taken into consideration.

For example, previous studies only define the study of Euclidean and spatial network variants to determine the new path for a given set of safe zones, to find paths that minimize the distance travelled outside the safe zones. The main aim of the study was to strictly avoid the entire region with the exception of the safe zones.

To make the safest path more refined we have included few more attributes like we have devised a transformation of continuous data which provide the information about the density of the people in the entire area and the frequency of automobiles. Our module also includes the number of police stations, malls and hospitals, which are present in the entire path of travel. This research methodology follows Dijkstra’s Algorithm, which can efficiently produce the shortest path for selection of the route. This technique also claims to be among the best approaches in solving the shortest path problem and always provide the shortest path from one node to another. To make the result more accurate and flawless we will take note of cellular strength and criminal activities happening in the different locations. These help users to know about all the vulnerable zones if present in the path of travel and alert them before selecting the path.

III. METHODOLOGY

This project is focused on analyzing all the potential indicators from a stock so that maximum amount of information can be gathered to understand a stock’s behavior. Information about the reactivity of the stock towards the market, knowledge about its correlation with other stocks which makes the process of creating a portfolio easier. We can generate a report for every stock to analyze whether it is a stock with higher daily

© 2018 JETIR October 2018, Volume 5, Issue 10 www.jetir.org (ISSN-2349-5162)
returns or a stock which is good for long-term investment. Indicators like momentum and Bollinger Bands gives us an idea about the current situation of a stock whether it’s a buy or sell.

Module for Portfolio optimization is also integrated in which using the Sharpe ratio as indicator is used to give the best possible allocation to invest among a series of stocks to generate the maximum profit. This is achieved by using SLSQP algorithm ‘Sequential Least Squares Quadratic Programming’ in this it checks the gradient descent of the function line to provide the best-optimized results by keeping all the bounds and constraints in mind.

The model used is an ensemble which is a model constructed from aggregating multiple machine learning models to make more accurate predictions. We have collected all the major potential indicators like Sharpe Ratio, Bollinger Signals, Simple Moving Average, Momentum of the stock Volume change etc., and all these indicators are the features of machine learning model. The labels i.e. buy, sell or hold are then mapped while training the dataset using models like Linear regression, K nearest Neighbor, Random forests, and SVM all these models are then passed through a Voting classifier in which all these models provide their predictions and then voting is done the one with the best results during the testing phase is selected for the predictions to give the best accuracy\textsuperscript{7].

After training this model is used for predicting the values of stocks after 7 days from the current day.

![Methodology Diagram](image1)

**Figure 1: methodology**

**IV. SYSTEM ARCHITECTURE**

The system architecture is something which defines the behavior and more the view of the system. The system architecture of our approach is shown in figure 2 which displays the various modules possessed in the system. The module’s description is described below:

![System Architecture Diagram](image2)

**Figure 2: system architecture**
4.1 Data extraction and pre-processing
First, we collected historical data of S&P 500 companies from Quandl. Then the data is converted into CSV file to read and view it. After getting the data we used python library to manipulate the data to produce useful information which is then used as input. Panda’s library provides the function which is used to remove the NaN and Zero values from the dataframe. There is column available in dataframe called index which helps to access rows. There is column available in dataframe called index which helps to access rows. Slicing is done to take adjusted close price of each stock. All the values taken are adjusted according to the stock-splits if any so that there are no haywire values to disrupt are prediction mode

4.2 Computation of Potential Indicators
Multiple modules of analyzing stocks are combined to predict if the following day’s closing price would increase or decrease. All five methods needed to be in agreement with the algorithm to predict a stock price increase or decrease. The modules are Bollinger Bands (BB), Momentum (M), Sharpe ratio (SR), Correlation(C)\(^9\).

4.2.2 Bollinger Bands
Bollinger bands are the indicators which provide buy and sell signals. It is based on a simple moving average. The graph is plotted as a grouping of 3 bands. The upper band is two standard deviations above a moving average; the lower band is two standard deviations below that moving average, and the middle band is the moving average itself\(^9\).

\[
\text{stdDev} = \sum_{i=1}^{N}(\text{price (i)} - \text{MA(N)})^2
\]

\[
\text{Upperband} = \text{MA} + \text{D} \sqrt{\sum_{i=1}^{N}((\text{Prices (i)} - \text{MA})^2)/N}
\]

\[
\text{Lowerband} = \text{MA} - \text{D} \sqrt{\sum_{i=1}^{N}((\text{Prices (i)} - \text{MA})^2)/N}
\]

D=No. of standard deviation

4.2.3 Sharpe ratio
Sharpe ratio is a method used for calculating the risk-adjusted return. It is a measure of excess portfolio return. A portfolio with a higher Sharpe ratio is considered superior relative to its peers. The Sharpe ratio is a ratio of return versus risk\(^10\).

\[
\text{Sharpe ratio} = \frac{c_p - cf}{\text{Portfolio standard deviation}}
\]

Where
- \(c_p\) = Expected portfolio return
- cf = Risk free rate
- stdp = Portfolio standard deviation

4.2.4 Momentum
The Momentum indicator designed to identify the speed (or strength) of price movement. It compares where the current price is in relation to where the price was in the past.

\[
\text{Momentum[t]} = \frac{\text{Prices[t]} - \text{Prices[t-n]}}{\text{Prices[t-n]}}
\]

t= Current day

n=Period of days

4.2.5 Correlation
The correlation coefficient is a measure that determines a mutual relationship or connection between two or more things. The range of values for the correlation coefficient is -1.0 to 1.0. If the correlation is greater than 1.0 or less than -1.0, a mistake has been made. If the value is -1 then it is a negative correlation, while the value 1.0 correlation is a positive correlation.

\[
\rho_{xy} = \frac{\text{Cov}(r_x, r_y)}{\text{std}_x \times \text{std}_y}
\]

In this module keeping Sharpe ratio as our main indicator, we optimized the allocation of the stocks which are present in the user’s portfolio. The allocations are selected in such a way that it gives the maximum amount of return and optimizes the Sharpe ratio to maximum This is achieved by SLSQP algorithm ‘Sequential Least Squares Quadratic Programming’. The algorithm asks for an initial guess to start the process and then from that point it checks the gradient descent at that point and then moves in a particular direction until it gets a local maxima after obtaining the local maxima then it moves in the opposite direction until it gets a maxima and then both maxima are compared to obtain a global maxima which is our optimized portfolio with revised allocation\(^10\).

4.3 Portfolio optimization
In this module keeping Sharpe ratio as our main indicator, we optimized the allocation of the stocks which are present in the user’s portfolio. The allocations are selected in such a way that it gives the maximum amount of return and optimizes the Sharpe ratio to maximum This is achieved by SLSQP algorithm ‘Sequential Least Squares Quadratic Programming’. The algorithm asks for an initial guess to start the process and then from that point it checks the gradient descent at that point and then moves in a particular direction until it gets a local maxima after obtaining the local maxima then it moves in the opposite direction until it gets a maxima and then both maxima are compared to obtain a global maxima which is our optimized portfolio with revised allocation\(^10\).

4.4 Training Dataset
All the potential indicators which are calculated during the Analysis phase are the main features during the training of the dataset. The whole dataset used for prediction is of the past decade we have divided the dataset into 4 equal parts for training and testing. During training of the dataset, the features are mapped with labels to give accurate predictions. Labels are the targets of our predictions like a percentage change in price or a momentum which is base of our prediction model as we are using the ensemble approach we are using multiple machine learning models for prediction\(^11\)\(^12\).

Algorithms used for Training the Dataset are as follows:

© 2018 JETIR October 2018, Volume 5, Issue 10 www.jetir.org (ISSN-2349-5162)
4.4.1 K Nearest Neighbour

This algorithm is used for classification in which K, a kernel is used which is a number by default either 3 or 5 which can be changed according to the need, the value of K decides how many nearest neighbors are checked are to be referred before generating a prediction value.

4.4.2 Linear Regression

Linear regression is one of the most used machine learning algorithms in which a line is drawn using the dataset which has a slope and intercepts, according to the dataset the line can be of any degree to make it best for better predictions[13].

4.4.3 Random Forests

In this multiple random trees are generated the more the trees the lesser are the chances to overfit the dataset and in this finding the root node and splitting the feature nodes will run randomly. This algorithm is good for both classification and regression purposes.

4.5 Testing Dataset

As the dataset was divided into 4 equal parts it is tested in the shuffle with other 2 parts and cross-validated. All the above-mentioned algorithms are passed through a voting classifier in which all the algorithms predictions are evaluated and voted to get the best accuracy for future predictions.

4.6 Prediction using Ensemble

By using an ensemble approach we get a very high accuracy on our dataset without overfitting the dataset. The advantage of this approach is that the algorithm selected benefit each other like KNN is good at classification, Linear regression is good at regression and Random-forest is good at both so they provide us a very balanced result.

V. RESULTS

All the results which are generated in the graphical images and portfolio optimizations are from 26th January 2014 to 26th January 2018 only the Prediction result of the prediction model is from 2004 to 2018.

5.1 General Analysis

Initially, general Analysis is done on a single stock like Amazon ‘AMZN’ the figure shows the details
The graph is divided into Three Parts, the first part of the graph shows the open high low close of the stock, the second part shows the volume change and the 3rd part is the Bollinger band of Amazon stock.

![Graph](image)

**Figure 4: daily returns**

This graph shows the daily returns of Apple, Google and Amazon all together in different colors to show the which among them is more volatile or good for daily stock marketing and from the graph it can be easily deduced that Amazon is better among these 3 stocks as it has more spikes. Mean daily return of these 3 stocks is Apple with .09%, google with .08% and Amazon with .16%

![Graph](image)

**Figure 5: scatterplot google vs amazon**

Scatter Plot shows the correlation between two stocks meaning if one rises then what are the chances that other will rise too. The correlation value of Apple and Google is .755 meaning if Apple goes 1% up then there is a chance that Google will go .75% up
This graph shows these two stocks which will give the higher return on a daily basis as Amazon bar is higher so Amazon will give a higher return.

5.2 Portfolio Optimisation

For our Experimental Setup, we took 5 stocks
Amazon Google IBM Apple and Abbott Laboratories the Allocation initially were taken 10%, 20%, 20%, 30% and 20% respectively and total amount spent on the portfolio was 10000 US dollars.
Give Portfolio
AMZN GOOGL AAPL IBM ABT
Give alloc
.r .2 .3 .3 .3
Optimization terminated successfully. (Exit mode 0)
Current function value: -2.7819419969610283 Iterations: 8 Function evaluations: 57 Gradient evaluations: 8
Investment amount 10000
total return before optimization 8764.307616091843
Total return after Optimization 22462.299971916247
Sharpe Index before Optimization 1.9277036987125546
Sharpe ratio after Optimization 2.78104
[57.05218029 1. 30.62270911 1. 10.3629265]

Figure 9: portfolio optimisation

Our Algorithm Returned
If we followed the given allocation we would have made 8764$ meaning we would have been at a loss but if we use the allocations Given by the algorithm i.e. 57%, 1%, 30%, 1% and 10% then we would make 22000$ which is a significant increase.

5.3 Prediction Analysis

In this the data is from 2004 to 2018 around 12 years in this we predicted the price is Apple stock from 23rd March onwards with an Accuracy rate of 75.3%

VI. CONCLUSION AND FUTURE WORK

The result shows that this algorithm produces accurate and reliable predictions which give consumer better solution where to invest their money and produce more return. An attempt has been made to build an efficient predictive model of the stock market where the trend for the next day is predicted. By considering various patterns like continuous up/down, volume traded per day and also including reactivity of each stock towards the market and correlation among each stock. The model has been built and tested with S&P 500 stock market data available open source. The input data used in the model is the preprocessed historical closing values of the index. The model predicts the next trading days closing value of the index. The training dataset comprises of the historical data previous to the prediction day of prediction. On the considered dataset, Random forest, KNN is performed to predict prices[14].

As this model completely comprises of technical analysis in future a machine learning model can be created which comprises of fundamental analysis also i.e. about the news on stock market and companies how it will affect the market because it is an important factor which cannot be ignored at all by doing we can predict the prices more accurately and efficiently[15].

References


