Stock Market Analysis and Prediction.

Rahul Myneni
Univesity of Texas, Arlington
Rahul.myneni@mavs.uta.edu

Pavan Vinnakota
Univesity of Texas, Arlington
saivenkatapavan.vinnakota@mavs.uta.edu

Abstract

Stock market prediction is the process of trying to estimate the future value of a company stock or other financial institution traded on a financial stock exchange. The successful prediction of a stock's future price will maximize gains and minimize the losses. This project proposes a machine learning model to predict stock market price. Proposed model is based on the study of stocks historical data and technical indicators. The algorithms select best free parameters combination for SVM and LSTM to avoid over-fitting and local minima problems and improve prediction accuracy. The proposed model was applied and evaluated using ten benchmark financials datasets and compared with artificial neural network. The obtained results showed that the proposed model has better prediction accuracy.

Our study shows that the Neural Networks outperform Support Vector Machines on average which implies that the existence of temporal dependencies in financial time series.

Index Terms- Machine Learning, Time Series forecasting, Neural Networks, Financial Analysis

Introduction:

In the financial world, stock trading is one of the most important activities. This is a process of buying and selling the stocks. Professional traders have used a variety of analysis methods such as fundamental analysis, technical analysis and quantitative analysis for the past several years. These analytical methods use different sources ranging from news to price data, but they all aim at predicting the company's future stock prices but they are not so effective. For the past few years, the increasing knowledge of machine learning in various fields have enlightened many people to implement machine learning algorithms to predict, and some of them have produced quite promising results. In this project, we will focus on short-term price prediction on general stock using time series data of stock price.

Artificial Neural networks and support vector machines are two of the most useful tools to implement any time series forecasting problems so we also used these algorithms. But while using neural networks there might be a problem of overfitting which can be avoided by using the input data the right way.

Dataset Description:

We used the Yahoo Finance datasets of the 10 of the top technological companies like Apple, Microsoft etc. for the past 10 years. This has information of the volume of stocks, opening price, closing price, highest, lowest price of the day as we are using short-term prediction this will be sufficient. But as we know the stock prices are based on the trends of the price we need to consider the momentum of the stock and the index for the past week and also the change based the previous days price.

Dataset Preprocessing:

The raw data that was available has many problems. Some of the data was missing in a few rows, not all the features have much effect on the accuracy and can cause over/under fitting. To avoid these problems the following steps have been applied on the dataset.

1.Feature Dropping: Features with high density of missing values repeatedly are dropped.

2.Feature Extraction: The features that show maximum dependency on the accuracy are extracted by checking the change in accuracy when each feature is dropped.

3.Trend Summarization:

Our target is to obtain values for short term, while many features from the raw datasets show clear trend over time.

These can affect the ability to predict the right way. Therefore, we take the percent change of the prices in consecutive days.

4. Regularization (Feature Scaling):

As the scales of the features vary drastically depending on the stock we used regularization limit the values to a particular range to improve the performance of the stock models.

Selecting the time frame:

We used the data from Q1 2010 to Q4 2017 which has 1700 entries for each stock.

5.Creating Features from existing data:

All the 12 stocks belong to the same sector so the sector momentum has been calculated, stock and index momentums and volatility have been considered.

Index momentum is the average momentum of the index for the past n days.

Index volatility is the average percent change over the last few days.

After preprocessing we ended up with 8 features and 12 stocks. Then we divided the dataset into 75-25 Training and testing.

Flow of project:

Step 1:
Feature Extraction and selection.
Step 2:
Optimizing and training SVM and LSTM
Step 3:
Testing LSTM and SVM models with new data
Step 4:
Computing Loss and accuracy and mean square error

Algorithms Implemented:

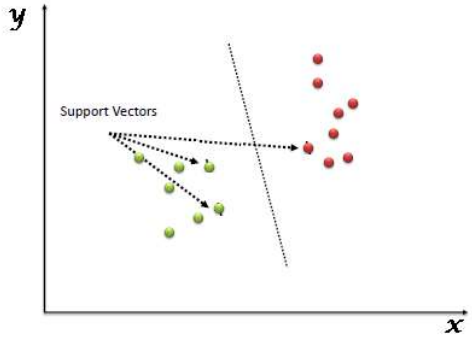
SVM:

SVM also known as Support Vector Machines is a discriminative classifier defined by a separating hyperplane. Given the labeled data the algorithm produces an optimal hyperplane for categorizing new data. SVM model represents the examples as points in space mapped so that the examples of the separate categories are divided by a clear gap which is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

SVM perform both linear and non-linear classifications, the non-linear classifications are performed by Kernel Trick. Kernel trick implicitly maps the inputs into high-dimensional feature spaces.

SVM also works when the data is not labelled i.e. where the outputs are not known, they use Support-Vector Clustering to perform clustering of data to groups, this is widely used.

Maximal-Margin Classifier is a hypothetical classifier that uses the Maximal-Margin hyperplane, the margin is the distance between the line and the closest data points on either side, the optimal line which can separate the two classes is the largest margin. The points which are closer to



the line are called support vectors which define the hyperplane. The optimization procedure followed here is the one which maximizes the margin.

As the real data can be mixed up we can separate it with a hyperplane, hence we use soft margin classifier which relaxes the constraint of maximizing the margin of the line resulting in some points of training data violating the line.

SVM Kernels are used to practically implement SVM algorithm.

Linear Kernel:

Here we take the dot-product as the distance measure from new data to support vectors.

kernel K(x, xi) = sum(x * xi)

Polynomial Kernel:

Here we use a polynomial instead of a dot product

Kernel K(x,xi) = 1 + sum(x * xi)^d

If d =1 the it is same as the linear kernel

Radial Kernel:

This is the complex kernel ,

Kernel K(x,xi) = exp(-gamma * sum((x – xi^2))

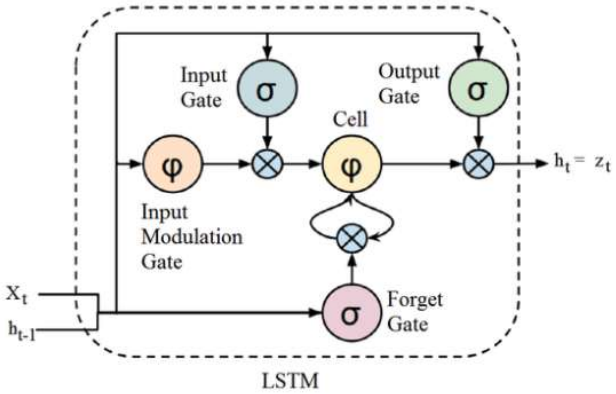
Gamma often lies between 0 and 1.

Radial Kernel is used in the project because it outperforms polynomial kernel and linear kernel cannot be used because stock markets trends are not linear.

Long Short Term Memory(LSTM):

LSTM is a deep learning technique developed to tackle with vanishing gradients problem that occurs in long sequence. LSTM has a cell, 3 gates output, forget and update gates. The cell remembers value over arbitrary time. These gates are used to regulate the flow of information in and out of the cell and also the information that needed to be used as activation for the next layer. The output gates help us determine the amount of information to output as activations to the next layer.[3]

Using LSTM we can store the values of changes of the past



few days and use it for the prediction of the next days value. Using LSTM, adding the features such as closing value, stock momentum, volatility, index momentum, index volatility and sector momentum and storing the past weeks values in the memory cell.

Sequential model has been implemented in the project

LSTM has been implemented using various hyperparameters, loss functions and optimizers.

The layers have been added and the neural network has been built accordingly to maximize the accurarcy.

Dense layers, activation functions all have been accordingly adjusted.

Loss functions: Mean Squared error, Mean Absolute errors etc.

Optimizers: Adagrad, SGD, adam have been used.

Comparative research works:

Many research papers implemented the algorithms using the data such as opening, closing, volume, date and implemented Support Vector Machines and LSTM [1]

Difference In approach:

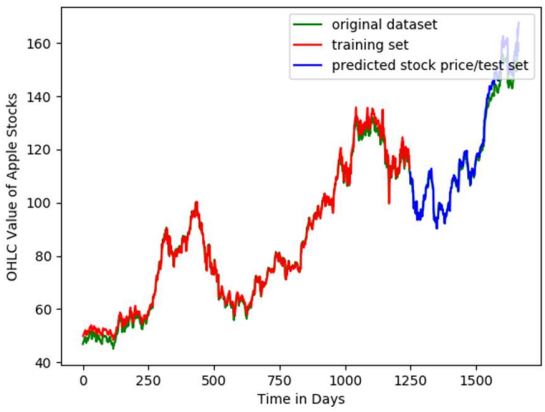
Most of the projects use the open, close, high values to analyze the stock markets but that is not very effective. So, we used the stock, index, sectors changes and momentum which helped us achieve better results.

Results:

By using LSTM on stock markets we were able to achieve the accuracy of 67% with a standard deviation of 2 percent which is much better than the accuracies in the research papers that we referred to on S&P500 data.

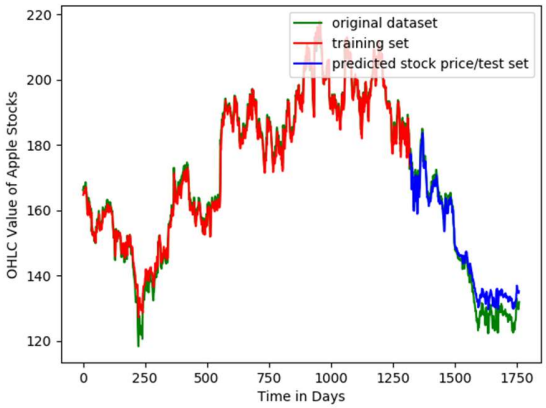
Whereas by using SVM on stock markets a accuracy of 65 with a standard deviation of 1.84% has been achieved.

Apple stock prediction using adagrad and lstm




```
Epoch 1/5
 - 3s - loss: 0.0042
Epoch 2/5
 - 2s - loss: 3.7515e-04
Epoch 3/5
 - 2s - loss: 3.1348e-04
Epoch 4/5
 - 2s - loss: 2.6510e-04
Epoch 5/5
 - 2s - loss: 2.3218e-04
Train RMSE: 1.86
Test RMSE: 3.23
Next Day Value: 168.649658203125
```

IBM Prediction:




```
Train RMSE: 2.36
Test RMSE: 4.04
Next Day Value: 136.21255493164062
```

Contributions: Both the team members were involved in all the stages of the project. The background research and analysis of related work was shared evenly. The processing of data set and selection of features has been done by Pavan. The implementation of SVM and LSTM by Rahul. The results were analyzed by both.

6. Analysis:

1.What did we do well?

We did a very good job in converting the dataset available to make sure that we consider the trends of the stock, index and also the top 10 technology stocks so that the effects of change of price of 1 stock may alter the other.

2. What can still be done?

Further improvements can be made by analyzing the reports and news related that might effect the stocks.
Along with this research, the analysis of how economic growth model will affect in stock market prediction in comparison to the neural network models and with specialized machine learning techniques.

3. What could have I done better?

We could have analyzed the news and extract information that might effect the stock price but it requires a completely different set of skills in natural language processing.

Conclusion:

Concluding from our results, we were only able to achieve an accuracy of 70 in the best case by considering the past data and the technical indicators.
We also observed that the accuracy is very high when the price fluctuation is low on average of 2% of the stock price. But there are some anomalies because some times there might come out an excellent product which drastically improve the stock prices. Also, the results show that the values predicted are very close to the given value for most cases. There is a chance that the stock price has hit the particular price on that day but cannot prove it because the data only has the opening high low, close values and the mean of the values has been considered for evaluation of the algorithm.
Though the accuracy is not very high there is very high chance that the stock hit that price on that particular day.

References:

[1]Alice Zheng, Jack Jin Using AI to Make Predictions on Stock Market.

[2]SVM:https://en.m.wikipedia.org/wiki/Support-vector_machine

[3]https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[4] LSTM theory:https://en.wikipedia.org/wiki/Long_short-term_memory

[5] Osman Hegazy 1, Omar S. Soliman 2 and Mustafa Abdul Salam3 2013 A Machine Learning Model for Stock Market Prediction