Group-7
Data Mining assignment

Literature survey

Anshul Agarwal IWM2017008, Arya Krishnan ICM2017501
Harsh Aryan ITM2017003, Priyal Gupta IHM2017004
Ritik Raj Gupta IIM2017003

Paper-1: Expert System with Applications - Stock market one-day ahead movement prediction using disparate data sources.

Citation-
Weng, Bin & Ahmed, Mohamed & Megahed, Fadel. (2017). Stock Market One-Day Ahead Movement Prediction Using Disparate Data Sources. Expert Systems with Applications. 79. 10.1016/j.eswa.2017.02.041.

Problem addressed-

Predictions of early stock were focused on the **Efficient Market Hypothesis(EMH) and random walk theory.** Some early models indicated that stock prices can not be predicted because they are driven by news instead of present/past values. So current news could predict the wrong results. Earlier their accuracy was less than 50%.
This prediction problem can be divided into two parts:
(1) Knowledge-base
(2) AI algorithms.

The Technique to solve the problem-

This paper blends disparate online data sources with conventional time series and technical indicators for a stock that can provide more accurate results.
Diversifying the knowledge base by combining data from disparate sources such as:

1: The number of Wikipedia page visits along with Wikipedia Momentum, Google Relative Strength Index etc.
2: Different technical indicators to predict the stock market value.
3: Google’s news data that comprises daily news about a company's products or other related things.
4:Stock market information, including price opening/closing, NASDAQ and DJIA indices, and so on.

The selection of features is based on Outcomes(hereafter targets) and Recursive feature elimination (RFE) to select features to predict better results.

This paper compares ANN, SVM, and DT(Decision tree) for prediction.

Dataset used-

Dataset is taken from Yahoo finance website:

1. Open price and close price.
2. High or low, and volume of the stock
3. Movements in DJIA and NASDAQ composite indices
4. Price to earnings ratio(P/E) as an assessment of the company's condition.

Dataset is collected from 1/05/2012 to 1/06/2015 i.e. 37 months, with a hit ratio of 85%.

Results-

So, finally, the best predictive results are obtained when disparate data sources are included with the other factors. Wikipedia provides more information than google news. Thus the combined form of all factors provides a prediction with better accuracy.
Thus, the model predicts best for Target 2.
Target 2 : Open(i + 1) − Open(i).
It compares the next day's opening price with today’s opening price.
This paper improved the accuracy by more than 20 % by using SVM and decision trees along with disparate data sources.

Comparison of the results -

The accuracy rate on stock market value prediction-

Algorithm	Accuracy Rate
Previous:	
SVM(with old technical indicators)	65%
This paper:	

SVM(with old technical indicators with other 3 factors) 85%

Paper-2: Sentiment and Knowledge Based Algorithmic Trading with Deep Reinforcement Learning

Citation-
Nan, Abhishek & Perumal, Anandh & Zaïane, Osmar. (2020). Sentiment and Knowledge Based Algorithmic Trading with Deep Reinforcement Learning. arXiv:2001.09403v1

Problem addressed-

The lack of reliable labeled data that considers all the factors that dictate the ups and downs of the market, has made it difficult for supervised learning attempts to give dependable predictions. The supervised Learning models used aren't able to correctly model the nature of the market. There are too many variables involved in the real world which makes it almost impossible to have reliable algorithms for automated stock trading.

The technique to solve the problem-

The paper focuses on the use of sentiment analysis done on news related to a traded company and its services together with a reinforcement algorithm called Q-learning to learn an appropriate optimum policy to trade stocks of the given company. Q-learning is a model-free reinforcement learning algorithm. Given an environment, the agent tries to learn a policy that maximizes the total reward it gets from the environment at the end of an episode (a sequence of interactions).

To find the relevant news title on-which to apply sentiment analysis, the paper uses a traversal of a knowledge graph. It uses the knowledge graphs for exploiting news about implicit relationships. That is, for example, if Microsoft's Stocks' sentiment analysis is to be done, we find all the news articles related to Microsoft, Windows, Azure and related terms through the knowledge graph.

Dataset used-

Stock data: Stock data from the Yahoo Finance API dated from January 1, 2014, to December 31, 2017, was used for the training environment. The data for the test period is from January 1, 2018, to December 31, 2018.
Microsoft Corporation’s (MSFT) stock data - i.e. was used, the agent was trained to trade Microsoft stocks.

Sentiment Data: For news information, historical news headlines from the Reuters Twitter account was scrapped using a python scraper. The time period of the news headlines corresponds exactly to the stock data.

Results-

We compare both approaches, i.e. an agent with sentiment data provided and another agent without any sentiment data provided. The agent with no sentiment input does learn a policy good enough to make a profit, but nowhere near good enough as compared to the agent which had sentiment input.

We also evaluate the model using the Sharpe Ratio, a measure that is often used in trading as a means of evaluating the risk-adjusted return on investment. In modern portfolio theory, a Sharpe ratio of 1 is considered decent. About 2 or higher is very good and 3 is considered excellent.

The agent without sentiment data learns a pretty poor policy as well (albeit still better than the random policy or even no trade policy or passive policy), despite making profits whereas the agent which learned a trading policy along with the sentiment data procures the highest profits and also its decision making was very good as evidenced by its Sharpe ratio of 2.4 for MSFT, 2.2 for AMZN, and close to 2 for TSLA.

Comparison of the results with other models-

Sharpe Ratios for different approaches:

Agent	**Sharpe Ratio MSFT**	**Sharpe Ratio AMZN**	**Sharpe Ratio TSLA**
Random Policy	-2.249	-1.894	-2.113
Without Sentiment	-1.357	1.487	0.926

With Sentiment **(Proposed Model)**	2.432	2.212	1.874

Paper-3: Ensemble Learning Approach for Enhanced Stock Prediction

Citation-
S. Mehta, P. Rana, S. Singh, A. Sharma and P. Agarwal, "Ensemble Learning Approach for Enhanced Stock Prediction," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-5.

Problem addressed-

Recently, researchers have tried to improve the accuracy of stock market prediction using machine learning classifiers.
However, the relationship between the prediction factors and the prices is very complex and may not be implementable by individual classifiers. Hence, the use of an ensemble has been proposed in this paper. This paper has used multiple models and taken a weighted average of each model's opinions so as to get an aggregated prediction.

The technique to solve the problem-

The goal here is to unite the predictions of multiple models to make an averaged, less noisy, more robust and better accuracy prediction.
Here, the ensemble used consists of three base learners, an SVR (Support Vector Regression), an LSTM (Long short-term memory network) and multiple regression.
This proposed ensemble approach will improve the robustness of the model towards errors and also improves the accuracy by averaging and hence removing the noise better than single-model classifiers.

Dataset used-

Yahoo Stock (collected from April 1996 to April 2016) has been used as a dataset for training the three base learners as well as the ensemble.
The dataset collected consists of 8 features (date, Value, open, high, low, close, volume, adj close) and 5039 samples.

Results-

On the basis of the weighted average method, the weight values in base learners are assigned by accuracy. Multiple regression- 3, Support vector regression- 2 and long short term memory network- 1 i.e. least weight as it gives the least accurate result.
The accuracy rates are multiplied by the weights derived for the base learners and added to obtain the aggregate accuracy of the ensemble.

Comparison of the results with other models-

The accuracy rates obtained are as follows-

Algorithm	**Accuracy rate**
Multiple regression	**99.02**
SVR	**98.56**
LSTM	**97.63**
Ensemble proposed	**99.12**

Paper-4: Stock trading decisions using ensemble-based forecasting models: a study of the Indian stock market.

Citation -
Jothimani, Dhanya & Yadav, Surendra. (2019). Stock trading decisions using ensemble-based forecasting models: a study of the Indian stock market. Journal of Banking and Financial Technology.

Problem addressed -

Prediction of Non-Linear and non-stationary characteristics of financial series, mainly stock price and index is a very difficult task. That may be due to the diverse stock market dynamics ranging from effects of Financial news to macroeconomic conditions to investor's behavior.

All statistical and computationally smart models were used to forecast stock prices. Basically Auto Regressive Moving Average (ARMA) and Auto-Regressive Integrated Moving Average (ARIMA) assumptions of statistical models such as linearity, stationarity, and normal data distribution limit the modeling capabilities of these techniques for non-stationary and non-linear financial data. While non-linear data can be modeled using Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and its extensions, they do not capture the financial market data irregularities.

The Technique to solve the problem -

The idea of ensemble models is used in this paper to overcome the limitations of various statistical and computationally smart models. It operates on the premise that using several predictors increases precision in forecasting. The two techniques to represent ensemble models have been discussed: competitive and co-operative.

This paper talks about how the intended task is carried out in various parts such as :

1. It uses a two-phase ensemble framework to predict the stock prices comprising of -

   Phase 1: Empirical Mode Decomposition(EMD), Ensemble Empirical Mode Decomposition (EEMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) are the non-classical models implemented for the decomposition of historical stock prices to a set of subseries basically six relatively stationary IMFs(Intrinsic Mode Function) and a residual component. limitations on EMD's mode-mixing problems were resolved by the usage of EEMD and CEEMDAN.

   Phase 2: Artificial Neural Network and Support Vector Regression (SVR) are the machine learning algorithms implemented in this phase.
   Here, each subseries is forecasted using these algorithms. Firstly Lag Parameter is calculated using the decomposed sub-series which is used to calculate the input values of SVR and ANN models.
   These inputs are then used for carrying out the prediction part. Both of the models are supervised algorithms for machine learning. Hence, the first 70 percent of data is used to train the model, and the remaining 30 percent is used to check the model.

   Finally, individual subseries forecasts are integrated in order to achieve the overall predictions.

3. The method of Root Mean Square Error (RMSE) is then used to evaluate predictor efficiency.

4. Statistical Analysis of the findings was done using Wilcoxon Signed Rank Test (WSRT), Friedman-Test and post hoc Nemeyi test.

5. Trading laws were also described in order to determine the best decisions to exchange stocks.

Dataset used -

The proposed structure is tested for a span of 8 years on constituents in the Nifty index where financial series considered for this analysis includes weekly close prices of Nifty index constituents from January 2008 through December 2015.

Nifty comprises 50 stocks covering 22 sectors making it a stronger Indian stock market representation than that of Sensex which contains only 30 stocks.

Results -

EMD-ANN, EEMD-ANN, CEEMDAN-ANN, EMD-SVR, EEMD-SVR, and CEEMDAN-SVR are the six hybrid models introduced utilizing constituents of Nifty index.

Different statistical tests are conducted on the models to compare their performance, such as a WSRT, Friedman-Test, and Nemeyi post hoc test.

Comparison of the results-

1. Hybrid SVR based on EMD and hybrid ANN based on EMD surpassed conventional SVR and ANN models.

2. Among the EMD-based hybrid SVR models CEEMDAN-SVR had better performance than EEMD-SVR, EMD-SVR and SVR models.

3. CEEMDAN-ANN efficiency among the EMD-based hybrid ANN models was found to be higher than the EEMD-ANN, EMD-ANN and ANN models.

4. The figures show that in all three ensembles, SVR (EMD-SVR, EEMD-SVR, CEEMDAN-SVR) performed higher than ANN (EMD-ANN, EEMD-CEEMDAN-ANN). The remaining models were outperformed by the CEEMDAN-SVR model.

5. Compared to the conventional Buy-and-Hold strategy, trade rules based on ensemble models yielded better ROI (return on investment).

Paper-5: Stock Market Index Prediction Using Artificial Neural Network

Citation-

Moghaddam M.H., Esfandyari M., Moghaddam A.H., “Stock market index prediction using artificial neural network”, Journal of Economics, Finance and Administrative Science, 21 (41) (2016), pp. 89-93.

Problem addressed-

Artificial neural networks also are known as ANN, have a variety of applications in the field of science and technology. Also, the recent discoveries show that artificial neural networks have a better ability to predict than statistical methods have because there are complex relations between different input variables.
This paper predicts the efficiency of ANN for predicting the daily NASDAQ stock exchange rate.

The technique to solve the problem-

This paper uses feed-forward ANN trained by the backpropagation algorithm. This paper takes into account the historical stock prices for a short period and also the weekdays as input. This can be represented by the equation

$$y(k) = f\,(y(k - 1), y(k - 2), y(k - 3), ..., y(k - n), D(k))$$

where y(k) is the stock price at time k, n is the number of historical days, and D(k) is the day of the week.
The model is built, trained and tested using MATLAB software R2010a. The determination coefficient (R^2) and the mean square error (MSE) were used to calculate the performance of the modeled output. R^2 is calculated as follows:

$$R^2 = 1 - \frac{\sum \left(y_{\mathrm{exp.}} - y_{pred.}\right)^2}{\sum \left(y_{\mathrm{exp.}} - \bar{y}\right)^2}$$

MSE is the square mean of variance between the expected values and the real values. The following equation is estimated to MSE:

$$MSE = \frac{\sum (y_{pred.} - y_{\exp.})^2}{M}$$

where y_{exp}. and y_{pred}. are the real and predicted values, respectively, and the total number of data is M, respectively.

Dataset used-

A robust model was created using the data of stock exchange rates of NASDAQ on a daily basis, dated from January 28, 2015, to 18 June 2015. The first 70 days (January 28 to March 7) were chosen as a training dataset, and the last 29 days were used to test the predictive potential of the model.

Results-

Applying the OSS training method and TANSIG transfer feature in a network with 20-40-20 neurons in hidden layers for four preceding working days resulted in an optimized trained network with R2 values of 0.9408 for the validation data set.
For nine preceding working days, the optimized network with validation R2 of 0.9622 is a network with 20-40-20 neurons in hidden layers OSS training method and LOGSIG transfer feature.

Comparison of the results-

The prediction ability of BPNN (backpropagation neural network) observed with different combinations of training functions and transfer functions:-

Training Function	Transfer Function	Accuracy
OSS	TANSIG	0.9267
OSS	LOGSIG	0.9069

Paper-6: Stock Market Prediction Using Machine Learning

Citation-

Chouhan, Lokesh & Agarwal, Navanshu & Parmar, Ishita & Saxena, Sheirsh & Arora, Ridam & Gupta, Shikhin & Dhiman, Himanshu. (2018). Stock Market Prediction Using Machine Learning. 10.1109/ICSCCC.2018.8703332.

Problem addressed-

Making predictions based on the values of the current stock market indices by training on their previous values with better accuracy than the previously used models.

The technique to solve the problem-

Regression and LSTM models are two models used in this paper. Regression involves minimizing error and LSTM contributes to remembering the data and results in the long run. The paper utilizes the gradient descent linear regression algorithm for predicting correct values by minimizing the error function. Linear Regression is performed on the data and then the relevant predictions are made. The R-square confidence test was used to determine the confidence score.

Dataset used-

In this project, supervised machine learning is employed on a dataset obtained from **Yahoo Finance**. The dataset consisted of approximately 9 lakh records of the required stock prices and other relevant values. The data reflected the stock prices at certain time intervals for each day of the year. **The test set was kept as 20% of the available dataset.** This dataset comprises of following five variables:

1. Open
2. Close
3. High
4. Low
5. Volume
6. Date
7.Symbol

Results-

LSTM and Linear Regression-based models have better accuracy than deep neural network techniques used before to predict the market.
These models can provide even better results with a greater percentage of data set.

The more the system is trained and the greater the size of the dataset utilized the greater the accuracy which will be attained. (This paper used 20% of the data set).
The LSTM model offered more accuracy than the Regression-based Model.

Comparison of the results with other models-

Regression-Based Model Results:
The R-square confidence test resulted in a confidence score of 0.86625.
LSTM Based Model Results:
The model resulted in a Train Score of 0.00106 and a Test Score of 0.00875.

TABLE-

Paper no.	Author	Problem addressed	Basic approach	Achievement	Limitations
1	Weng et. al.	Previous experiments were based on the EMH and random walk theory analysis of the stock market. Hence, accuracy was not good enough.	Combining disparate online data sources with traditional technical indicators to get better accuracy.	This paper improved the accuracy by more than 20 % by using SVM and decision trees along with disparate data sources.	Collecting all the data from Wikipedia, Google News, stock data and other technical indicators.
2	Nan et. al.	Algorithmic trading, due to its inherent nature, is a difficult problem to tackle, due to too many variables involved and lack of properly labeled data that can model the trend/nature of the market correctly.	The use of sentiment analysis done on news related to a traded company and its services together with deep Q - learning to get an optimum policy for trading.	The Sentiment + Reinforcement Learning model offered better profits than different models without Sentiment analysis. It also gives a better Sharpe ratio, which means a lesser risk.	Model buys/sells only one stock instead of “x” stocks. The model uses only open/ close price of a day, rather than close price and the volume of the last ”x” -days
3	S. Mehta et. al.	The target function may not be implementable by single model classifiers but may be approximated by the averaging of multiple models.	Using an ensemble approach to get an aggregated opinion of multiple models in prediction.	The deviation between the actual and predicted price has significantly reduced in the ensemble model.	Model diversity and deciding optimal weights for base learners are challenges.

4	Jothimani et. al.	Statistical and computationally sophisticated models do not detect financial-market data anomalies.	Paper has applied the concept of ensemble models.	The figures show that on all three ensemble models, SVR (CEEMDAN-SVR) performed better than ANN.TCompared to the conventional Buy-and-Hold strategy, trade rules focused on ensemble models yielded better ROI.	EMD is suffering from mixing of Mode. EEMD puts in a slight amount of noise. When using the boosting and bagging algorithms, the models can be further optimized.
5	Moghaddam et. al.	This paper predicts the efficiency of ANN for predicting the daily NASDAQ stock exchange rate.	Feedforward ANN which is trained using the backpropagation algorithm is used. Short-term historical stock prices and also the week-days are used as input.	In a network with 20-40-20 neurons in hidden layers, the OSS training method and TANSIG transfer feature resulted in an optimized trained network with high precision values.	Predicting the correct initial weights for the neural network can be challenging.
6	Ishita Parmar et. al.	Prediction making for the values of the stock market indices by training on their previous values to get better accuracy than the previously used models.	Regression aims to minimize the errors and LSTM to remember the data and results in the long run.	The LSTM model got more accuracy than the Regression based Model. It is concluded that the percentage of test sets used will affect accuracy.	Deciding the percentage of the data set which is going to be used as the testing data.