Market Price Prediction for Ethiopian Commodity Market

Selam Damtew[a], Eliyas Girma Mohammed[b], Surafel Lemma Abebe[b]

[a]School of Electical and Mechanical Engineering, Addis Ababa Science and Technology Universty, Addis Ababa, Ethiopia
solbdu@gmail.com
[b]School of Electrical and Computer Engineeing, Addis Ababa Institute of Technology, Addis Ababa Universty, Addis Ababa, Ethiopia
beliyas4@gmail.com, surafel.lemma@aait.edu.et

ABSTRACT

The current Ethiopian market is conducted in a traditional manner and market drivers are still not used for prediction of future market price. Although, large amount of market data have been gathered for a number of years by both governmental and non-governmental organizations, little have been done to analyze the data for future market price prediction. Moreover, the analysis methods were often manual creating inefficiency in time and quality of market prediction. Analyzing valuable data will show us what the future holds and accelerate the development goals of the country in the sector.

This study examines features of current Ethiopian market attributes to find out most valuable features for predicting market price. Eighteen technical indicators are taken and tested for their individual ability of prediction and redundancy using feature selection algorithm. The result shows that *Stochastic %K, Stochastic %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton* and *Moving Average Convergence/ divergence (MACD)* features are in the top ten and have a better prediction performance.

This study also compares the performance of *Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (K-NN) and Ensemble Learning* algorithms in predicting Ethiopian commodity market price. The outcome of feature selection was used to build models and compare the algorithms performance. We conducted experiment using separate train and test data. From the four algorithms *ANN* and *Ensemble Learning* are found to perform better than *SVM* and *K-NN*. The average *Mean Absolute Error (MAE)* rate of the *ANN* model was 2.8084. *Ensemble Learning* and *SVM* follow with average MAE rate of 4.9362 and 8.1178 respectively. The performance of *K-NN* model is the least with MAE rate above 45.3381.

KEYWORDS

Prediction; Machine Learning; Market Price Prediction

1 Introduction

Market prediction using different analysis techniques is regularly practiced in modern marketing systems by collecting and analyzing different market information. Traders in any part of the world are interested in a market that is profitable and use different technical indicators, macroeconomic factors and stock market indexes to study the market. These numerous market drivers provide information that reflects the existing market price characteristics and facilitates prediction of future market price. As a result we can prevent anticipated negative changes on the market due to new information of the market. Market analysis, however, is not a common practice in Ethiopia. A direct implementation of other countries' study findings on market prediction for the case of Ethiopia is impractical. Studies on other countries follow different approaches based on the countries' economical and market situation. Most of the studies (Tsai and Wang, 2009; Khan et al., 2011; Adebiyi et al., 2012; Das and Padhy, 2012; Shen et al., 2012; Karazmodeh et al., 2013; Soni and Shrivastava, 2010; and Narayanan and Govindaraian, 2015) we reviewed focus on stock market which is not introduced and experienced in Ethiopia. Moreover, the studies indicate that market features that have impact on one country may not have similar impact for another country. As a result, we need to take a closer look of the target market to form market strategies.

This paper explores the predictability of future Ethiopian market price enabling governmental and non-governmental organizations in the sectors, and individual trades to conduct market activities with minimum business risks. Recently, Ethiopian Commodity Exchange (ECX) (www.ecx.com.et) started hosting commodity market and disseminates market information on coffee, beans, sesame and grains on real-time basis, offering contracts for further delivery. Although such systems can be appreciated, sufficient data analytical activities are not employed to predict future market price scenarios resulting in market uncertainty. Traders are still seeking market analysis which indicate future opportunities and reduce business risk. To explore the information in accurate way, we can encode the data in to technical indicators and be able to predict market prices using machine learning algorithms. Machine learning algorithms can be used to find patterns in data and predict future market price of goods in Ethiopian market. The study aims to identify market features which influence the prediction of Ethiopian market price.

Twenty two features were evaluated for their predictive ability using the model *Relief attribute evaluator* (Robnik-Sikonja and Kononenko, 1997). We found that *Stochastic %K, Stochastic %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton* and *Moving Average Convergence/ divergence (MACD)* are the top ten features that have better predictive ability. These features are used as an input for the machine leaning algorithms. In this study, we compared four commonly used machine learning algorithms, i.e., Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (K-NN) and Ensemble Learning, to identify the suitable algorithm in Ethiopian context.

The result shows that the most efficient algorithms in predicting market price in Ethiopia are ANN and Ensemble Learning.

2 Case study

2.1 Data

The collected data covered the time period of January 2008 up to January 2015. The collected data from ECX contained features namely: *Trade date; lowest price (low), highest price (high), volume (ton), opening* and *closing price* of pea bean, coffee and sesame. These six features are used to compute 16 technical indicators that are used as additional features of the market.

The data from ECX contained 94,993 records in which the majority of the records are of coffee, which were around 72,160. Sesame and pea beans have 18,021 and 4,812 records respectively. Table 1 shows the original 6 features and the computed 16 technical indicators.

2.2 Data preprocessing

The dataset was checked for completeness and correctness of the required features and integrity before conducting the analysis and prediction. We considered a data to be missing if the values for the weekdays are not recorded. Such missing values are observed in three of the original datasets from ECX (coffee, sesame and pea been). To overcome the effect of missing values, we filled the missing values using interpolation technique. We used the *Trade Date* column to locate the missing days from the data. We create two data frames; one containing the original dataset and the other contains full dates of the year. For every day there needs to be at least one record of data. However, for a single day there could be many records. For every missing day, we insert a row containing NA values which represents missing value. After inserting a row and populating the row with NA values, we interpolated the missing values. The interpolation is done by considering the neighboring values and the overall values for that specific date of all the years. The Linear interpolation is given by:

$$y = y_1 + \frac{(x-x_1)(y_2-y_1)}{x_2-x_1} \quad (1)$$

Table 1.The initial 22 features that are used for feature selection

No	Feature	Description
1	Trade date	Date of trade
2	Closing price	The final price the commodity is sold with
3	High	The highest price given by bidders
4	Low	The lowest price given by bidders
5	Ton	Volume of the commodity provided for bid
6	Opening price	Opining price stated by the bidder
7	EMA	Exponential Moving Average
8	Close gain/loss	The gain or loss from previous day market
9	RSI	Relative Strength Index
10	SMA-20	Simple Moving Average of 20 days
11	BB-Upper	Upper Bollinger Bands
12	BB-Lower	Lower Bollinger Bands
13	MACD Fast	Moving Average Convergence/divergence Fast
14	MADC Slow	Moving Average Convergence/divergence Slow
15	MADC	Moving Average Convergence/divergence
16	MADC Signal	Moving Average Convergence/divergence Signal
17	Highest high	The highest high over the look up period
18	Lowest low	The lowest low over the look up period
19	Stochastic %K	Calculated with other quantity %D
20	Stochastic %D	Simple moving average of %K
21	20-days mean deviation	20 day mean deviation

2.3 Feature selection algorithm

The feature selection activity ranks features based on their individual ability in prediction using model *Relief Feature Attribute Evaluator* (Robnik-Sikonja and Kononenko, 1997). The Relief method is instance based learning. Having the training data δ, sample size m and a threshold of relevance τ, the model detects those features that are statistically relevant to the target concept. The relevant threshold is between 0 and 1 ($0 \le t \le 1$). The Relief is valid when the selected threshold can preserve relevant features and discard irrelevant features. Relevant features have a higher level of relevance and the reverse is true for irrelevant feature.

The model works for nominal and numeric feature scales. The difference of two feature values X and Y is given by

$$\text{diff}(X_K, Y_k) = \begin{cases} 0 < if\ X_K \text{ and } Y_k \text{ are the same} > \\ 1 < if\ X_K \text{ and } Y_k \text{ are different} > \end{cases} \quad (2)$$

When X_K and Y_k are nominal,

$$\text{diff}(X_K, Y_k) = (X_K - Y_k)/\text{nuk} \quad (3)$$

When X_k and Y_k are numerical,

Where, nuk is a normalization unit to normalize the values of diff into the interval [0, 1]

The model takes a sample composed of m triplets of instances X, its Near-hit instance and Near-miss instances. Both the Near-hit and Near-miss instance are found in close neighborhood of X. The Near-hit, however, is found in the same category of X, while the Near-miss is not found in the same category. The Near-hit and Near-miss selection is based on p-dimensional Euclid distance. Features with average weight relevance above the given threshold τ will be selected.

2.4 Comparison of machine learning algorithms

The machine learning algorithms selected for this study are SVM, K-NN, ANN and Ensemble Learning. The algorithms are selected based on their advantage and past performance seen in other research.

SVM is selected due to the following reasons (Huang, 2005); (1) Data classification could be performed without making strong assumptions; (2) SVM is established on the structural risk minimization principle, which seeks to minimize an upper bound of generalization error, and is shown to be very resistant to the

Table 2. Features in order of importance, from higher to lower

	Features
Pea Bean	%K, %D, Close Gain/Loss, Closing Price, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA, SMA_20, MACD Slow, BB-Upper, BB-Lower, Highest High, Lowest Low, MACD Signal, Trade Date
Sesame	%K, %D, Close Gain/Loss, Closing Price, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA, SMA_20, MACD Slow, BB-Upper, BB-Lower, Highest High, Lowest Low, MACD Signal, Trade Date
Coffee	%K, %D, Closing Price, Close Gain/Loss, High, Opining Price, Low, RSI, Ton, MACD, MACD Fast, 20-Days Mean Deviation, EMA, SMA_20, SMA, MACD Slow, BB-Upper, MACD Signal, Highest High, Lowest Low, BB-Lower, Trade Date

over-fitting problem; and, (3) SVM model is a linearly constrained quadratic program so that the solution of SVM is always globally optimal, while other models may tend to fall into a local optimal solution.

ANN is included in this work because (Tsai and Wang, 2009); (1) As a component of ANN fails, the net continues to operate (based on its highly parallel nature), and; (2) A neural network learns and does not have to be re-programmed.

K-NN is selected because (Teixeira and Oliveira, 2010); (1) The cost of learning process is zero, (2) Learning does not require making any assumption about the characteristics of the concepts, and; (3) Complex concepts can be learned by local approximation using simple procedures.

Ensemble Learning is included in this work because (Narayanan and Govindarajan, 2015); (1) Ensemble learning combines predictions from multiple models and, hence, the results are more diversified; and, (2) Gives more robust estimate of a statistical quantity with a low bias and a high variance.

The performance of each machine learning algorithm was tested using features extracted from the dataset in feature selection stage. Then based on prediction results, best performing algorithms are selected. To test the algorithms the dataset is divided in to two. The first partition was used for training, while the remaining is used for testing. The test dataset contains a data of different year that is not used for training. The training data set covers the time span of 2008-2014 and the 2015-2016 first quarters for testing.

The prediction performance is evaluated using mean absolute error (MAE), and root mean squared error (RMSE). MAE measures the average magnitude of the errors in a set of prediction, without considering the direction. It expresses average model prediction error in units of the variable of interest. The smaller the values of MAE, the closer are the predicted values to the actual values. RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It is square root of the average of squared differences between predicted and actual value.

3 Results and discussions

3.1 Feature selection

The first activity associated with feature selection is testing each individual attributes contribution on the predicted prices. The feature selection tool used to see the performance of each attribute was ReliefFAttributeEval which is available in Weka tool. The results are shown in Table 2.

The result from ReliefFAttributeEval feature evaluator includes a number of features. The given 22 features from Table 1 were ranked based on their predictive ability. For the three datasets (pea bean, sesame and coffee) used in the study, a relatively similar results were recorded (see Table 2). Features %K, %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and MACD are found in top 10 ranked list for all three datasets. The ranking shows that these features can explain the pattern in the data better than the rest of the features; and, hence, give better prediction accuracy.

3.2 Machine learning algorithm comparison

The experiments were done using a separate train and test set. From the given data of coffee and pea bean, we use 84% of the data for training and the other part which is 16% for test and for sesame 68% for training and the other part for test. All the training data set covers the time span of 2008-2014 which is indicated by 84% and 68% respectively. The tests set includes years from 2015-2016 first quarters which is 16% of coffee and pea bean and 32% for sesame. The percentage difference is due to the difference in the collected data. The test set contains a data from a different year than the year used to train the model. To build the machine learning models, we used the top 10 features ranked in the previous experiment. These features have better prediction ability, and reduce the computational cost and processing time needed for prediction, if all features were used.

The experiment results using separate train and test data for the four prediction models is shown in Table 3. The performance of ANN and Ensemble Learning was good as it showed lower

Table 3 Prediction Performance using Top Ten Features

Model		Coffee	Sesame	Pea bean	Average MAE
SVM	MAE	11.5981	9.9888	2.7667	8.1178
	RMSE	18.5323	14.2259	5.2227	
ANN	MAE	1.2166	5.6839	1.5248	**2.8084**
	RMSE	3.7112	8.561	5.7375	
K-NN	MAE	29.7984	33.7384	72.4777	45.3381
	RMSE	39.2275	47.2856	88.6036	
Ensemble Learning	MAE	3.3094	7.2889	4.2103	4.9362
	RMSE	4.9128	10.9571	6.3391	

Table 4 Prediction values for ANN and Ensemble Learning

	Days	Actual	Predicted for ANN	Predicted for Ensemble Learning
Pea bean	Day 1	825	824.89	829.39
	Day 2	760	759.91	756.08
	Day 3	885	885.09	880.45
	Day 4	858	848.23	854.42
	Day 5	945	944.99	945.44
Coffee	Day 1	1030	1024	1021
	Day 2	796	795	802
	Day 3	1400	1406	1402
	Day 4	1178	1176	1177
	Day 5	760	756	757
Sesame	Day 1	2700	2676.89	2680.5
	Day 2	2237	2251.83	2275.79
	Day 3	3370	3376.89	3370.45
	Day 4	3320	3314.24	3310.87
	Day 5	2670	2631.93	2645.49

average MAE. The models have recorded MAE of 2.8084 and 4.9362 respectively. Table 4 shows actual and predicted values for the models with minimum MAE, ANN and Ensemble learning.

For the results of the models in Table 3, we perform ANOVA test for the MAE and RMSE performance metrics. Table 5 shows the results from ANOVA test. The analysis of variance indicated that the mean absolute error of machine learning models, viz., ANN, Ensemble and SVM are not significantly different at α=0.05. Contrastingly, the mean absolute error of the K-NN prediction model was significantly higher as compared to the other models used in the study.

4 Conclusion

Despite the large market data collected in Ethiopia, market price prediction is conducted using a traditional approach. The traditional approaches are manual and, hence, are time consuming and less precise. In this study, we examined current Ethiopian market attributes to find out most valuable features for predicting market price. Eighteen technical indicators are computed from the collected and tested for their individual ability of prediction and redundancy using feature selection algorithm. The result shows that Stochastic %K, Stochastic %D, Close gain/loss, High, close price, Opening Price, Low, RSI, Ton and Moving Average Convergence/ divergence (MACD) features are in the top ten and have a better prediction performance.

Using the outcomes of feature selection, we compared the performance of four machine learning models (SVM, ANN, K-NN and Ensemble Learning) in terms of their price prediction accuracy. We compared models with separate train and test data using feature of individual predictive ability. From the four models, the performance of ANN and Ensemble Learning

Table 5 Mean Separation for Mean Absolute Error

Machine learning	MAE	RMSE
Ensemble	4.94[a]	7.4030[a]
ANN	2.81[a]	6.0032[a]
SVM	8.12[a]	12.6603[a]
KNN	45.34[b]	58.3722[b]
Mean	15.30	21.109675
Minimum Significant Difference @ α=0.05	36.794	42.048

Means with separate letters denote significantly different models.

algorithms were found to have better performance than SVM and K-NN. The average MAE rate of the ANN model was 2.8084. Ensemble Learning and SVM follow with average MAE rate of 4.9362 and 8.1178 respectively. The K-NN model was least performer with the MAE rate above 45.338.

ACKNOWLEDGMENTS

We would like to thank ECX and its staff for cooperating and providing the data required for this study.

REFERENCES

Tsai, C. F., & Wang, S. P. (2009). Stock price forecasting by hybrid machine learning techniques.In Proceedings of the International MultiConference of Engineers and Computer Scientists 1(755).

Das, S. P., &Padhy, S. (2012). Support vector machines for prediction of futures prices in Indian stock market. International Journal of Computer Applications, 41(3).

Khan, Z. H., Alin, T. S., &Hussain, M. A. (2011).Price prediction of share market using artificial neural network (ANN). International Journal of Computer Applications, 22(2), 42-47.

Adebiyi, A. A., Ayo, C. K., Adebiyi, M. O., &Otokiti, S. O. (2012).Stock price prediction using neural network with hybridized market indicators.Journal of Emerging Trends in Computing and Information Sciences, 3(1), 1-9.

Karazmodeh, M., Nasiri, S., &Hashemi, S. M. (2013). Stock price forecasting using support vector machines and improved particle swarm optimization. Journal of Automation and Control Engineering, 1(2), 173-176.

Soni, S., &Shrivastava, S. (2010).Classification of Indian stock market data using machine learning algorithms.International Journal on Computer Science and Engineering, 2(9), 2942-2946.

Narayanan, B., & Govindarajan, M. (2015).Prediction of Stock Market using Ensemble Model. International Journal of Computer Applications, 128(1), 18-21.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005).Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513-2522.

Teixeira, L. A., & De Oliveira, A. L. I. (2010).A method for automatic stock trading combining technical analysis and nearest neighbor classification.Expert systems with applications, 37(10), 6885-6890.

Shen, Shunrong, Haomiao Jiang, and Tongda Zhang.(2012)Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, CA, 1-5.

Robnik-Sikonja M., & Kononenko I. (1997), An adaptation of Relief for attribute estimation in regression. In 14th International Conference on Machine Learning, 296-304.