Stock Market Prediction Using Machine Learning Techniques

Mehak Usmani[1], Syed Hasan Adil[2], Kamran Raza[3] and Syed Saad Azhar Ali[4]

[1,2,3] Department of Computer Science
Iqra University
Karachi, Pakistan

[4] Department of Electrical Engineering
Universiti Teknologi Petronas
Bandar Seri Iskandar, Malaysia

[1] mehak.usmani@gmail.com, [2] hasan.adil@iqra.edu.pk, [3] kraza@iqra.edu.pk, [4] saad.azhar@petronas.com.my

***Abstract*-The main objective of this research is to predict the market performance of Karachi Stock Exchange (KSE) on day closing using different machine learning techniques. The prediction model uses different attributes as an input and predicts market as Positive & Negative. The attributes used in the model includes Oil rates, Gold & Silver rates, Interest rate, Foreign Exchange (FEX) rate, NEWS and social media feed. The old statistical techniques including Simple Moving Average (SMA) and Autoregressive Integrated Moving Average (ARIMA) are also used as input. The machine learning techniques including Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF) and Support Vector Machine (SVM) are compared. All these attributes are studied separately also. The algorithm MLP performed best as compared to other techniques. The oil rate attribute was found to be most relevant to market performance. The results suggest that performance of KSE-100 index can be predicted with machine learning techniques.**

Keywords-Stock Prediction; KSE-100 Index; Neural Networks; Support Vector Machine.

I. INTRODUCTION

Predicting the future has always been an adventurous and attractive task for the probing individuals. This kind of prediction becomes more fascinating when it involves money and risk like predicting the Stock Market. Research has been done on Stock market prediction by researchers of different fields including the business and computer science [1,2]. Researchers have tried different approaches for market prediction including different techniques & algorithm and different combination of attributes. The attribute that makes a prediction model depends upon the factors on which market performance can depend.

On the basis of the literature review, different attributes are selected for the prediction model. Most of the research work done in stock market is by the people related to business field. Although some research has also been made by computer science researchers but it still shows a vast area to be explored. In Computer Science the prediction may often relates to data mining or machine learning. Talking specifically about the Karachi Stock Exchange, there are very few researchers who have used the machine learning techniques for KSE market prediction. The main objective of this research is to find out whether the combination of different techniques that includes statistical, analytical and data mining techniques can predict stock market or not and up to what accuracy level.

As the main focus in this study is to predict the market, there exist few theories that are valid as well as oppose each other. The theory of random walk says price of a security can't be predicted using the historical data. It supports the argument that the difference between old price and current price of a security is completely independent. On the contrary, the chartist theories say there is some hidden information in the historical prices of a security that gives a clue to future price of that security. In [4], the researchers have used historical data to predict the position of stock market. The results of [4] proved that historical data has strong predictive ability. In [5], the research was performed on Asian markets to find out the factors that have concrete impact on market performance. In [6], the researchers have used Artificial Neural Networks (ANN) and statistical technique ARIMA on almost 3 year's data to predict KSE-100 index. The study presented in [7] has done a comprehensive analysis of the underlying relationship between macro-economic factors and KSE market. Similarly, few researches like [8] showed that person's mood plays the critical role in decision making. If the collective mood of the public is found using social media, then this can also help in predicting the decision they will make about investing the money in the market and thus market performance. In [9], the researchers proposed the use of data collected from different global financial markets with machine learning algorithms to predict the stock index movements. The study [10] made the use of financial NEWS in order to predict market prices. Current and historical NEWS about companies, economic and

 978-1-5090-2549-7/16/$31.00 ©2016 IEEE

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.

political events can help in stock prices prediction. Similarly, different statistical techniques have also been used like in [11] the researchers have applied the variable moving average (VMA) on data of Vietnamese Stock Market.

II. METHODOLOGY

In this research the stock market is tried to be predicted by three variants of Artificial Neural Network and by the Support Vector Machine algorithm.

A. Single Layer Perceptron

The first technique used for market prediction is Single Layer Perceptron model which is the most basic arrangement. The SLP contains an input layer and an output layer. The neurons in the output layer receive the weighted (*w*) sum of input neurons.

(1)

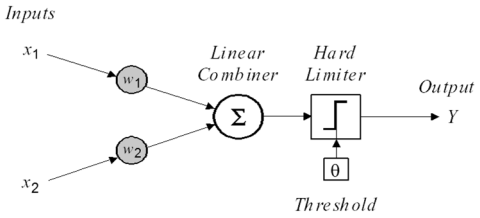

Figure 1. The Single Layer Perceptron Model

The weights in each iteration are calculated by the weight updating rule;

$$w_i\ (p+1)\ =\ w_i(p)+\ \Delta w_i(p) \quad (2)$$

$$\Delta w_i(p) =\ \alpha . X_i(p)\ .e(p) \quad (3)$$

Here, α is learning rate and e is error. The actual output is calculated by

$$Y(p) =\ Step\ [\ \textstyle\sum_{i=1}^{n}\ X_i(p)\ .w_i(p) -\ \theta\ ] \quad (4)$$

B. Multi-Layer Perceptron

The multi-layer Perceptron is a feed-forward neural network with one additional layer called the hidden layer. The hidden layer can be one or more. This layer also contains the intermediate neurons. Now the output neurons depend on output of hidden layer neuron and hidden layer neurons depend on input layer neurons for their processing. The intermediate output of the neurons in the hidden layer is calculated by;

$$Y_j(p) =\ Sigmoid\ \left[\textstyle\sum_{i=1}^{n}\ X_{ji}(p)\ .w_{ij}(p) - \theta_j\ \right] \quad (5)$$

Similarly, the output of neurons in output layer is calculated by

$$Y_k(p) =\ Sigmoid\ \left[\textstyle\sum_{i=1}^{m}\ X_{jk}(p)\ .w_{jk}(p) - \theta_k\ \right] \quad (6)$$

The weight update rule in output layer of MLP is

$$w_{jk}\ (p+1)\ =\ w_{jk}(p)+\ \Delta w_{jk}(p) \quad (7)$$

The weights in hidden layer will depend on the weights of output layer to calculate the error gradient δ and is calculated by,

$$\Delta w_{ij}(p) =\ \alpha . X_i(p). \delta_j(p) \quad (8)$$

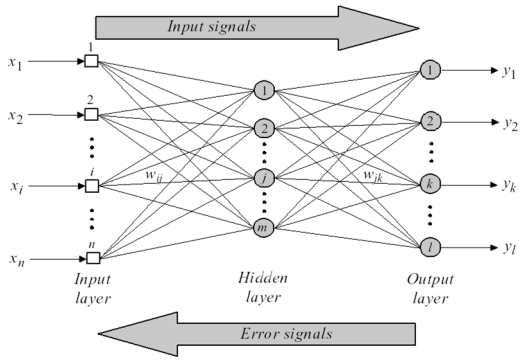

Figure 2. The Multi-Layer Perceptron Model

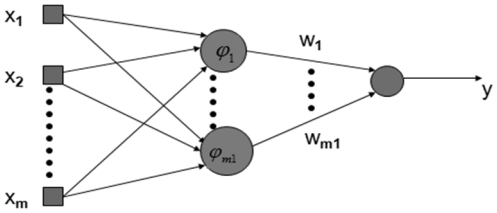

Figure 3. The Radial Basis Function Model

C. Radial Basis Function

This type of neural network is supervised algorithm and is a feed-forward network. It depends only on the radial distance from a point. Like MLP, the RBF [3] networks also have three different layers including input, output and hidden layer. Each hidden neuron represents a single radial basis function which has its center and width (spread). RBF transforms the non-linearly separable classes into the linearly separable classes.

The non-linear transfer function φ of RBF is given below. Here, the σ is a parameter called the spread, which indicates the selectivity of the neuron.

$$\varphi_{ij} = e^{-\frac{\left\| x - t_i \right\|^2}{2\sigma^2}} \quad (9)$$

Weights are computed by means of the pseudo-inverse method.

$$w_1\varphi_1(\|\ x_i - t_1\ \|) + \ldots + w_{m1}\varphi_{m1}(\|\ x_i - t_{m1}\ \|) = d_i \quad (10)$$

This can be re-written in matrix form as;

$$\begin{bmatrix} \varphi_1(\|\ x_1 - t_1\ \|) \ldots \varphi_{m1}(\|\ x_1 - t_{m1}\ \|) \\ \ldots \\ \varphi_1(\|\ x_N - t_1\ \|) \ldots \varphi_{m1}(\|\ x_N - t_{m1}\ \|) \end{bmatrix} [w_1 \ldots w_{m1}]^T = [d_1 \ldots d_N]^T \quad (11)$$

$$[w_1 \ldots w_{m1}]^T = \Phi^+ [d_1 \ldots d_N]^T \quad (12)$$

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.

D. *Support Vector Machines*

SVM is considered to be one of the most suitable algorithms available for time series prediction. This supervised algorithm can be used in both, regression and classification. The SVM involves plotting of data as point in the space of n-dimensions. These dimensions are attributes that are plotted on particular co-ordinates. SVM algorithm draws a boundary over the data set called the hyper-plane, which separates data into two classes as shown in the Fig. 4.

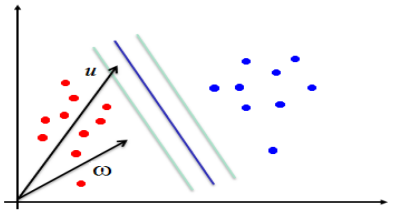

Figure 4. The Support Vector Machine Decision Boundary

The hyper-plane is the decision boundary which is later extended or maximized on either side between the data points. Considering the same figure, if u is some unknown data point and w is a vector which is perpendicular to the hyper-plane, then the SVM decision rule will be

$$\vec{\omega}.\vec{u} + b \geq 0 \quad (13)$$

The width of the hyper-plane must be maximized to increase the spread

$$W = [2/\|\omega\|] \quad (14)$$

$$W = max[2/\|\omega\|] \quad (15)$$

Applying Lagrange's Multiplier as

$$L = 0.5\,\|\omega\|^2 - \sum \alpha_i [y_i\,(\omega_i x_i + b) - 1] \quad (16)$$

$$L = \sum \alpha_i - 0.5\,\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \quad (17)$$

The updated Decision rule will be

$$(\sum \alpha_i\ y_i\ x_i)\,.u + b \geq 0 \quad (18)$$

III. The Proposed Method

The technique proposed in this research is to use different factors impacting the market as input attributes for the model. The output of the model is one of the two defined classes that are Positive Market and Negative Market. Fig 5 shows the structure of the proposed model. The model studies and compares the results of all four machine learning algorithms defined above. All the attributes that are used in this model were continuous numeric values and were of different range. These attributes are therefore normalized between [+1,-1] because all the parameters used can have positive and negative values. Each attribute will be discussed separately.

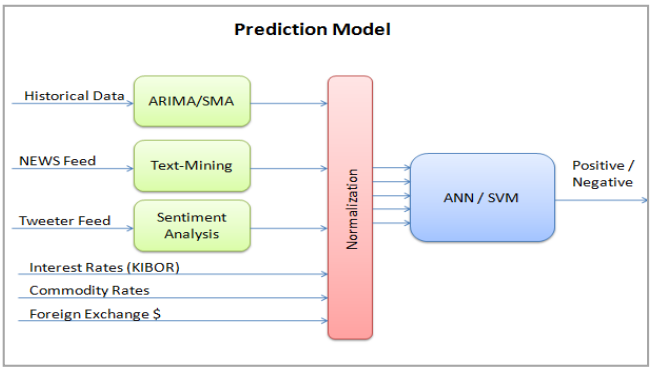

Figure 5. The Prediction Model

A. *Factors*

Following are different factors that were found to have some impact on market performance in different studies.

1. Market History
   The first attribute used as an input for the model is the historical closing index of KSE-100. The historical data was not made part of the model directly but after applying statistical techniques including ARIMA and SMA over the data. The window size used is 4.

2. The NEWS
   NEWS is another influencing factors considered for market performance. NEWS can be of different category but in this model only business, financial, political and international event based NEWS were included.

3. General Public Mood
   The market performance greatly depends on the investors' mood and sentiment. The collective sentiment of people may drive stock market performance. This can be achieved with the help of social media. In this research, twitter is used as a source of public sentiment.

4. Commodity Price
   The changes in prices of different commodities do have an impact over the market behavior. Change in price of commodities like petrol reflects on almost all items. Commodities including Gold, Silver and Petrol are used as input to the model.

5. Interest Rate
   The interest rates issued by State Bank of Pakistan to all the banks that provides loan to their customer also have effect on market. The Karachi Inter Bank Offer Rate (KIBOR) is issued on daily basis for different durations. In this study, 1-week rates are used.

6. Foreign Exchange
   Change in Foreign Exchange rate has been assumed to affect the market performance by many. Historical

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.

exchange rate between the Pakistan Rupee (PKR) and the US Dollar (USD) was used as an input to the model.

On basis of above factor, total of 9 parameters (i.e., Oil rates, Gold Rates, Silver Rates, FEX, SMA, ARIMA, KIBOR, NEWS, Twitter) are used as input for the prediction model, whereas the two classes are designed to be output by the mode.

B. Scope

The data used in this study spread over three months, from September, 2015 to January, 2016. The data gathered from NEWS and twitter lie between current day close till the next day's market opening.

C. Text Mining

NEWS and twitter data was available in the form of feed which was processed using text mining techniques. The library OpinionFinder [12] was used for this purpose. In twitter feed the use of non-English words was very frequent. This could cause the wrong classification of the text. The work was done to implement the dictionary that translates the Urdu word written in Roman to alternate English word. Later this updated feed was send to feed processor application that reads the whole text and classify it in one of the two classes.

IV. The Simulations & Results

The data was collected and developed so that it can be converted into the form that can be used in the model as inputs. The NEWS and twitter feed for each day was gathered and processed to be declared as Positive or Negative. Similarly, data from different sources was collected for the rest of the parameters. Because of the dependency on two parameters (NEWS and twitter feed) for current day's data, only 100 instances could be collected. The 100 records span over 4 months that is from September, 2015 to January, 2016. Out of these 100 instances, 70 were used in training and the remaining 30 were included in the test data set.

A. Implementing Machine Learning Algorithms

All four machine learning algorithms were separately applied on these data sets. The figures mentioned below show the actual and predicted market performance. The X-axis represents the instances of data set and Y-axis shows the classes. The Positive class is represented by 1 and Negative class is represented by 0. Fig. 6 shows the results of Single Layer Perceptron model trained by training set and tested on the different data set. The model gave about 60% accurate results. The model was unable to predict the Positive class well. When the SLP algorithm was tested on same data set that was used for training, it gave 83% accuracy as shown in Fig. 7. The SLP model's parameters were tuned to get more accurate results. The best results were achieved for 0.3 learning rate and 50 epochs.

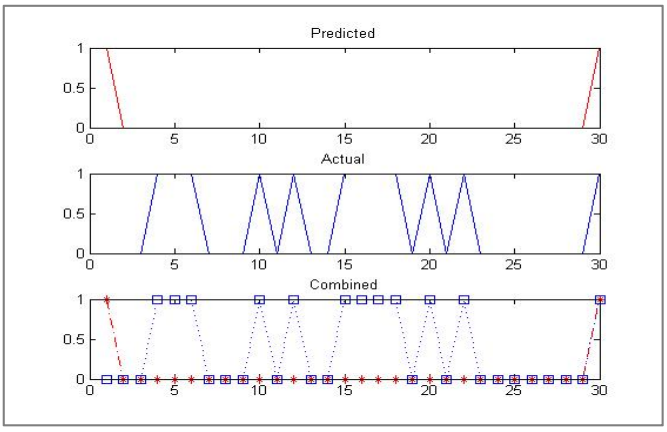

Figure 6. SLP Model verified on Test Data Set

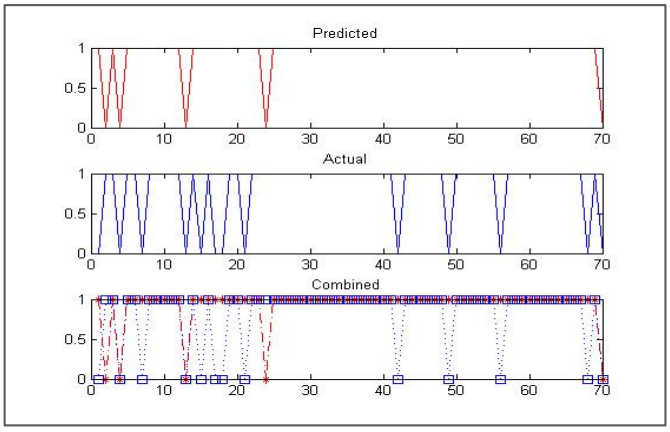

Figure 7. SLP Model verified on Training Data Set

Similarly, MLP algorithm was applied on data set first on training and then on testing. The model was also tuned for the best parameters. The most suitable parameters were found to be 0.05 learning rate, 1 hidden layer with 4 neurons and performed 500 epochs. The model gave 77% correct results when verified on test set and 67% on training set. The model surprisingly under performed for training set. The same can be seen in Fig. 8 and Fig. 9.

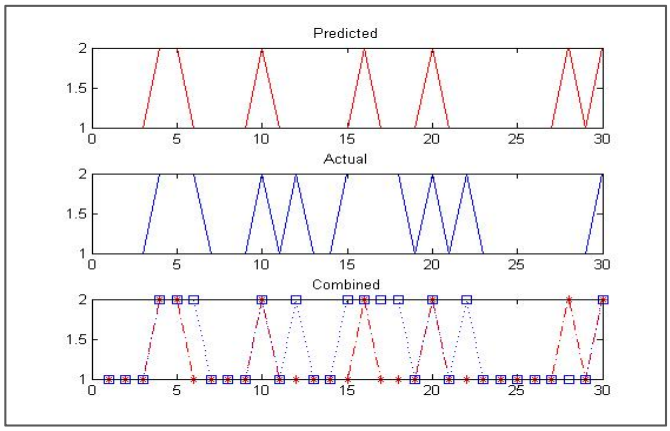

Figure 8. MLP Model verified on Test Data Set

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.

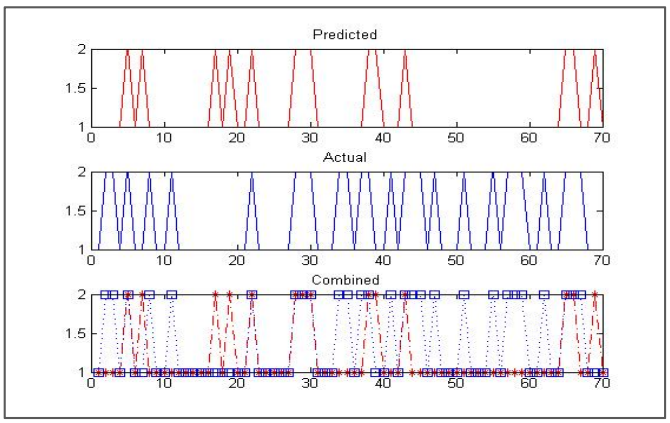

Figure 9. MLP Model verified on Training Data Set

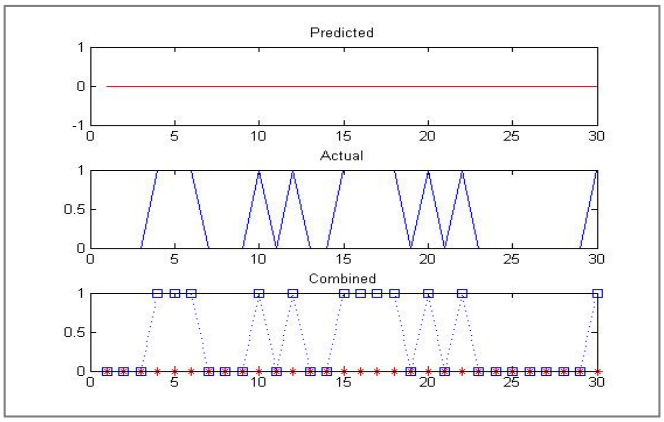

Figure 12. SVM Model verified on Test Data Set.

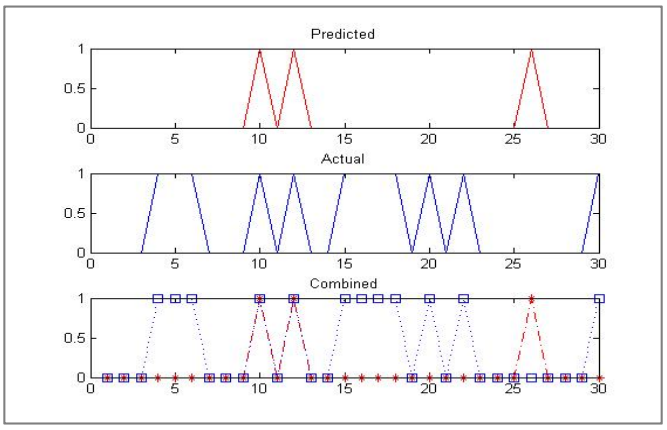

Figure 10. RBF Model verified on Test Data Set.

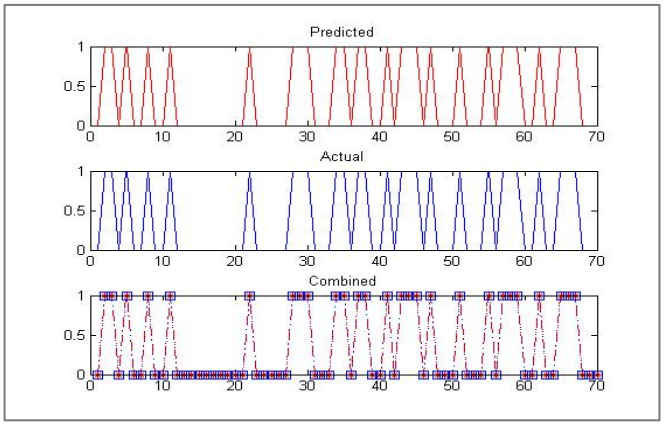

Figure 13. SVM Model verified on Training Data Set

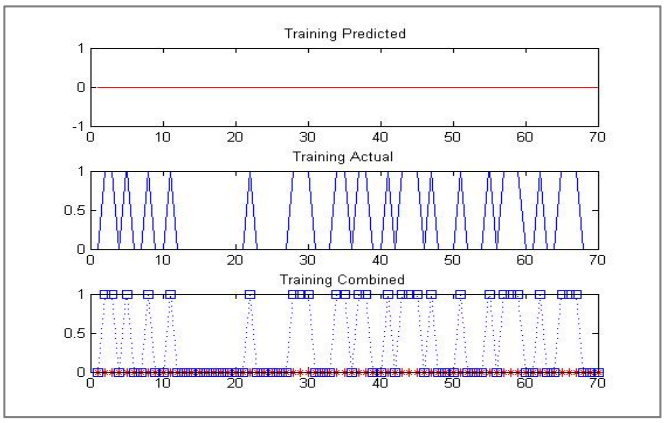

Figure 11. RBF Model verified on Training Data Set.

RBF algorithm was tuned to get maximum accuracy on prediction. The parameters used in RBF based model include Eta equals 0.03, 0.8 as interval and 500 epochs. The model gave 63% results when demonstrated on test set and 61% when demonstrated on training set as shown in Fig. 10 and Fig. 11.

Applying SVM algorithm on the training and test set gives different results. The algorithm produced 100% accuracy on training set but 60% on the test set. The kernel function used in the algorithm was RBF function and value of Sigma was taken as 0.3. The results can be seen in trailing figures.

The results of SVM algorithm was re-confirmed by changing the library used for classifying the instances according to SVM working but even that showed no improvement in the results.

B. Relationship between attributes & market performance

The factors expected to have some impact on market were used as attributes in the prediction model. The relationship between each of these attributes and market behavior was studied separately using Correlation and Co-variance. The attribute "Oil rates" showed co-variance of 422.5897 with market which is the largest values of co-variance among the attributes. The least impacting factor was found to be Foreign Exchange rate that showed value of 0.045 showing no association at all. The summarized result of all attributes finding correlation and co-variance can be seen in Table I.

The Fig. 14 shows the relationship between Oil Rate and market behavior and Fig. 15 shows the relation between Foreign Exchange Rate specifically Pakistan Rupee to US Dollar conversion rate. The figures show the strength of relationship between the attribute and the behavior of market. Foreign Exchange showed neither positive nor negative relationship which means this factor can be eliminated from the model as it does not have any contribution for class prediction.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COVARIANCE AND CORRELATION OF ATTRIBUTES & MARKET PERFORMANCE

S. No.	Attributes	Covariance	Correlation
1	Oil rates	422.5897	0.276812
2	Gold Rates	-586.7435	-0.07018
3	Silver Rates	4.9080036	0.0238194
4	FEX	0.045220645	0.00020486
5	SMA	-1500.61	-0.03858
6	ARIMA	-19741.62879	-0.131519799
7	KIBOR	-0.1984227	-0.0104532
8	NEWS	2.88594651	0.06580184
9	Twitter	-9.2570117	-0.1456551

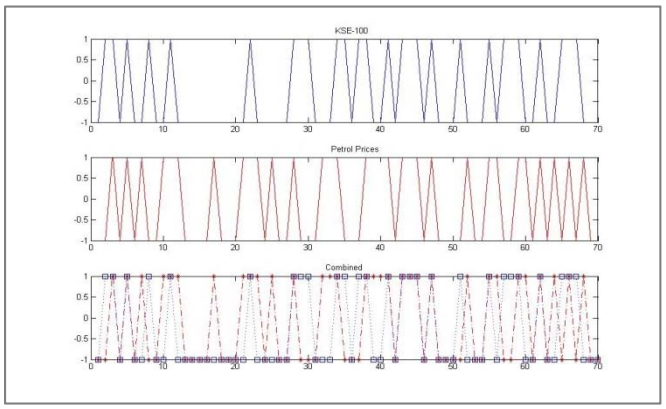

Figure 14. Relationship between Oil Rates and Market Performance

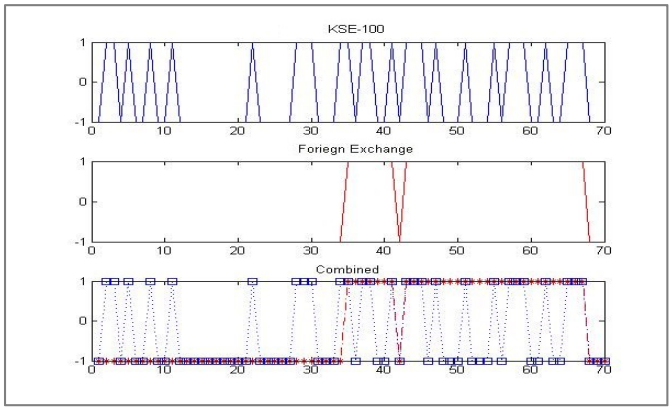

Figure 15. Relationship between FEX and Market Performance

C. *Performance Comparison of Algorithms*

The results of all four algorithms over training set and test set are compared in the Table II. It is evident from the comparison table, that SVM performed best on training set while the MLP algorithm did well on test data set. Conventionally, the prediction model works on test data set for the verification of the model that was built using the training data set. The test set has to be different from the instances used in the training set so that model can be tested on completely new and unseen instances. Therefore, MLP seems to be more efficient in predicting the market performance. While verifying the model's variants on test data set, MLP outperforms the remaining three algorithms.

TABLE II
COMPARISON OF MACHINE LEARNING TECHNIQUES

Data set used for verification	Machine Learning Algorithms			
	SLP	MLP	RBF	SVM
Training Set	83%	67%	61%	100%
Training Set	60%	77%	63%	60%
Average	71.5%	72%	62%	80%

V. CONCLUSION

The results of this study confirm that machine learning techniques are capable of predicting the stock market performance. Karachi Stock Market with KSE-100 index does follow a behavior that can be predicted using machine learning techniques. The Multi-Layer Perceptron algorithm of machine learning predicted 77% correct market performance. Even with the lack of resources and unavailability of data for the market, the model was able to predict the performance of the model to a good extent with only 100 instances that shows that KSE has the tendency to be predicted using machine learning techniques. Also the most related attribute to the KSE performance was found to be the petrol price and FEX proved to have no effect on the KSE performance.

REFERENCES

[1] S. Qamar, & S. H. Adil, "Comparative analysis of data mining techniques for financial data using parallel processing," In Proceedings of the 7th International Conference on Frontiers of Information Technology, 2009.

[2] S. H. Adil, & S. Qamar. "Implementation of association rule mining using CUDA," International Conference on Emerging Technologies (ICET), 2009.

[3] S. S. A. Ali, M. Moinuddin, K. Raza, & S. H. Adil, "An adaptive learning rate for RBFNN using time-domain feedback analysis," The Scientific World Journal, 2014.

[4] A. W. Lo, & A. C. MacKinlay, "Stock market prices do not follow random walks: Evidence from a simple specification test," Review of financial studies, vol. 1, no. 1, pp. 41-66, 1988.

[5] C. D. R. Aurangzeb, "Factors Affecting Performance of Stock Market: Evidence from South Asian Countries," International journal of academic research in business and social sciences, vol. 2, no. 9, 2012.

[6] S. Fatima, & G. Hussain, "Statistical models of KSE100 index using hybrid financial systems," Neurocomputing, vol. 71, no. 13, pp. 2742-2746, 2008.

[7] I. Ali, K. U. Rehman, A. K. Yilmaz, M. A. Khan, & H. Afzal, "Causal relationship between macro-economic indicators and stock exchange prices in Pakistan," African Journal of Business Management, vol. 4. No. 3, pp. 312, 2010.

[8] J. Bollen, H. Mao, & X. Zeng, "Twitter mood predicts the stock market. Journal of Computational Science," vol. 2, no. 1, pp. 1-8, 2011.

[9] S. Shen, H. Jiang, & T. Zhang, "Stock market forecasting using machine learning algorithms, 2012.

[10] S. Asur, & B. A. Huberman, "Predicting the future with social media," International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010.

[11] N. H. Hung, & Y. Zhaojun, "Profitability of Applying Simple Moving Average Trading Rules for the Vietnamese Stock Market," Journal of Business Management, vol. 2, no. 3, pp. 22-31, 2013.

[12] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, & S. Patwardhan, "OpinionFinder: A system for subjectivity analysis," In Proceedings of emnlp on interactive demonstrations, pp. 34-35, 2005.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on December 10,2022 at 01:33:58 UTC from IEEE Xplore. Restrictions apply.