International Journal of Computational Economics and Econometrics

ISSN online: 1757-1189 - ISSN print: 1757-1170
https://www.inderscience.com/ijcee

Improved stock price forecasting by streamlining indicators: an approach via feature selection and classification

Mohammad Javad Sheikhzadeh, Sajjad Rahmany

DOI: 10.1504/IJCEE.2023.10058606

Article History:

Received:	23 April 2022
Last revised:	24 June 2023
Accepted:	26 June 2023
Published online:	20 December 2023

Improved stock price forecasting by streamlining indicators: an approach via feature selection and classification

Mohammad Javad Sheikhzadeh and
Sajjad Rahmany*

Department of Computer Science,
School of Mathematics and Computer Science,
Damghan University,
Damghan, Iran
Email: sheikhzadeh@gonbad.ac.ir
Email: javadsheikhzadeh@yahoo.com
Email: s_rahmani@du.ac.ir
*Corresponding author

Abstract: Accurately predicting changes in stock prices is a complex and challenging task due to the multitude of factors influencing the stock market. Stock market analysts commonly rely on indicators for forecasting, but the interpretation of these indicators is often complicated and can result in inaccurate predictions. To enhance the precision of stock price forecasting, we propose a novel approach that incorporates feature selection algorithms and classification techniques. In fact, by identifying the most impactful indicators affecting each stock's price, the process of predicting stock prices will be significantly simplified. We conducted experimental tests on stock data from multiple companies listed in the Tehran Stock Exchange, spanning 2008 to 2021. Our findings demonstrate that reducing the number of features and indicators can significantly enhance the accuracy of stock price predictions in specific scenarios.

Keywords: stock forecasting; indicators; feature selection; classification.

Reference to this paper should be made as follows: Sheikhzadeh, M.J. and Rahmany, S. (2024 'Improved stock price forecasting by streamlining indicators: an approach via feature selection and classification', *Int. J. Computational Economics and Econometrics*, Vol. 14, No. 1, pp.42–60.

Biographical notes: Mohammad Javad Sheikhzadehn received his BS in Computer Science from Shahid Beheshti University and his MS in Computer Science from Tabriz University. He is a PhD student at Damghan University. Data mining, machine learning, wireless sensor networks, and image processing are among his research interests.

Sajjad Rahmany holds the position of an Associate Professor in the Department of Mathematics at Damghan University. Within this role, he imparts knowledge in various subjects such as machine learning, software,

and linear algebra. His academic pursuits encompass a strong focus on machine learning, deep learning, and their applications in other scientific domains.

1 Introduction

The stock market holds great significance as a financial market for investments, primarily due to its ability to generate profits in a relatively short period compared to other markets. Consequently, investors and traders are highly interested in predicting price fluctuations in the stock market. However, the unpredictable and turbulent nature of stock market data, characterised by nonlinearity, dynamism, instability, and chaos, poses a significant challenge in developing a reliable prediction system. Unlike conventional statistical data, stock market data exhibits intricate complexities, driven by erratic changes. Numerous factors, such as political, economic, seasonal, and market trends, influence the stock market (Chandar, 2020). To forecast stock market price changes, statistical models and machine learning models are commonly employed. While statistical models are straightforward and linear in nature, they struggle to uncover hidden insights within the nonlinear structure of stock market data (Chung and Shin, 2018). Therefore, the increasing application of artificial intelligence and machine learning methods has become prominent in addressing statistical problems, particularly in the field of financial market forecasting, including the stock market.

Several algorithms have been developed to predict price trends in financial markets. These algorithms include the back propagation neural network (BPNN), functional link artificial neural network (FLANN), wavelet neural network (WNN), recurrent neural network (RNN), radially bounded nearest neighbour (RBNN), grey wolf optimisation-Elman neural network (GWO-ENN), kernel principal component analysis (KPCA), support vector machine (SVM), and naive Bayes. Researchers have utilised these algorithms in various studies to forecast stock prices (Nahil and Lyhyaoui, 2018; Ray et al., 2014; Liao and Wang, 2010; Lei, 2018; Chandar, 2020; Rahimunnisa, 2019; Guo et al., 2013). However, due to the complex and nonlinear nature of market behaviour, it remains uncertain which algorithm performs best. Despite receiving less scientific attention compared to quantitative analysis (Murphy, 1999), a significant number of studies have employed technical indicators as the primary approach for forecasting stock prices.

Within the realm of financial market forecasting, there exist numerous indicators, yet some of them exhibit low accuracy despite the heavy computational load they require. As a result, it becomes imperative to identify a set of indicators that can effectively predict price changes in the market. One potential solution lies in feature selection algorithms, which can eliminate redundant features and reduce complexity, thereby enhancing the efficiency of forecasting methods. The studies conducted by Omuya et al. (2021) and Vandana and Chikkamannur (2021) have shed light on the effectiveness of feature selection algorithms in addressing this issue.

The utilisation of feature selection algorithms is based on three primary factors. Firstly, these algorithms can aid in mitigating the computational complexity associated with large datasets, which can pose difficulties in terms of cost-effectiveness and timely analysis. Secondly, feature selection algorithms have the potential to enhance the efficacy and comprehensibility of infrastructure processes. Finally, the incorporation of features that offer valuable insights can lead to an increase in classification power (Omuya et al., 2021; Ntakaris et al., 2020).

The primary challenge faced by stock market analysts is the presence of certain indicators that can lead to misjudgements about the trend of a particular stock. Improving the accuracy of forecasts requires the identification and use of key indicators for each stock, which necessitates feature selection. Feature selection plays an important role in enabling analysts to efficiently identify the most relevant and informative features from a large pool of potential candidates. This not only improves the accuracy of predictions but also simplifies and speeds up the computational process. However, working with high-dimensional data presents challenges as it may contain many irrelevant or uninformative features, which can complicate the computational process and result in storing excessive amounts of unnecessary information. In this research, multiple feature selection algorithms were used to reduce the number of indicators while three different classification methods, random forest, C4.5, and support vector machines (SVM) were utilised to evaluate the quality of the selected indicators. Our experimental findings revealed that the random forest classification algorithm outperformed other methods in several instances, demonstrating superior accuracy. Furthermore, by reducing the number of features and indicators, we achieved even higher prediction accuracy, particularly for weekly forecasts, while simultaneously reducing computational burden.

This paper is structured as follows. First, Section 2 provides a comprehensive discussion of related work. In Section 3, we describe the data and algorithms utilised in the study and present an overview of the performance evaluation methods employed. The experimental results are then presented in Section 4, followed by a thorough analysis of the findings in Section 5. Finally, Section 6 concludes the paper with a summary of our research and suggestions for future work.

2 Related work

There exists a significant body of literature exploring the use of machine learning models for predicting stock prices and trends. The majority of these articles demonstrate the substantial potential of such models in this domain. For instance, Lin et al. (2009) propose employing the PCA method for reducing and filtering noise in data, while Talebi et al. (2014) suggest a novel classification method for identifying upward, downward, and sideway trends in foreign exchange rates in the foreign exchange market. Similarly, Zhong and Enke (2017) investigate data mining methods for daily stock price forecasting and introduce various financial and economic features that are subsequently reduced using techniques such as KPCA and fast robust principal component analysis (FRPCA). Other researchers, such as Patel et al. (2015) and Dash and Dash (2016), predict stock price trends using technical indicators as input features for artificial neural network algorithms. Chen et al. (2018) compare the effectiveness of deep learning with machine learning methods such as back propagation (BP), extreme learning machine (ELM) and radial basis function (RBF) for predicting the behaviour of the Chinese stock market. According to Omuya et al. (2021), removing bands with no performance effect through the feature selection process is an effective means of reducing the need for performance. Obthong et al. (2020) explore machine learning algorithms and stock price prediction techniques. Xu et al. (2020) develop a stacked system for stock market prediction, leveraging the wavelet transform to reduce noise in market data and a stacked auto-encoder to filter out unimportant features from preprocessed data. As machine learning models become more complex, the interpretation of results and identification of patterns in data become increasingly challenging (Swales and Yoon, 1991). To overcome this challenge, feature selection and classification methods were employed in this study to enhance prediction accuracy and efficiency of stock price trends.

3 Proposed model

This study employed 25 attributes for each symbol, including open, close, high, low, volume, zig zag, RSI, Macd, signal, MFI, stochastic %D, stochastic %K, momentum, Kijun-Sen, Tenkan-Sen, Chikou-span, Senkou-span A, Senkou-span B, DEMA, DTOscSK, DTOscSD, NNRSI, OBV, MAOBV, and CCI (Talebi et al., 2014; Murphy, 1999, 2009; Marek and Cadkova, 2020; Patel, 2010; Di Lorenzo, 2012; Miner, 2008; Tsang and Chong, 2009). To ensure uniformity and avoid bias due to large values, all data was normalised and scaled to a range of [0, 1]. Multiple feature selection algorithms from the Weka library were used to reduce the number of indicators, including CFS subset evaluator, information gain attribute evaluator, gain ratio feature evaluator, correlation attribute evaluator, and symmetrical attribute evaluator (Bouckaert et al., 2016), while three different classification methods, random forest, C4.5, and SVM were utilised to evaluate the quality of the selected indicators. The 10-fold cross-validation method was conducted to validate the performance of these algorithms. A Java-based software program was developed and implemented to calculate the relevant indicators, apply feature selection and classification operations to the normalised data, and save the selected indicators and errors as separate files for each symbol. This allowed for systematic assessment of the performance of the different feature selection algorithms and classification methods on the data. The overall framework for the proposed model is illustrated in Figure 1.

The data used in this study consists of the daily trading history of three symbols (Shbhran, Kpars, and Valiz) from January 2008 to December 2021, comprising a total of 3,112 working days. This data was obtained from the Tehran Securities Exchange Technology Management Co website at http://tsetmc.com. In order to evaluate the prediction performance of the proposed model, we used the following performance measures:

- The mean absolute error (Chandar, 2020) is the average absolute difference between the predicted value and the actual value, and is calculated as follows:

$$MAE = \frac{1}{N}\sum_{k=1}^{n}|A_k - P_k|$$

- The root mean square error (Chandar, 2020) is the square root of the distance between the predicted and actual value:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{n}(|A_k - P_k|)^2}$$

- Classification accuracy is calculated by dividing the number of correct predictions made by the total number of predictions made and multiplying the result by 100 to obtain a percentage (Vandana and Chikkamannur, 2021):

  $$Accuracy = \frac{T}{N} \times 100$$

  where T is the number of samples correctly classified and N is the total number of samples.

- Sensitivity or recall indicates the percentage of the positive class that was correctly classified (Wong and Lim, 2011; Vandana and Chikkamannur, 2021):

  $$S_n = \frac{TP}{TP + FN}$$

- Specificity indicates how much of the negative class was correctly classified (Wong and Lim, 2011):

  $$S_p = \frac{TN}{TN + FP}$$

- The positive predictive value (PPV) is the probability of correctly predicting an outlier (Wong and Lim, 2011):

  $$PPV = \frac{TP}{TP + FP}$$

- The negative predictive value (NPV) is the probability that a non-outlier point was correctly predicted (Wong and Lim, 2011):

  $$NPV = \frac{TN}{TN + FN}$$

- F-measure or balanced F-score (F) is the harmonic mean of the PPV and sensitivity measures:

  $$F = \frac{2 \times PPV \times S_n}{S_n + PPV}$$

- AUC-ROC: the receiver operator characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against the FPR at various thresholds, essentially separating the ‘signal’ from the ‘noise’. The area under the curve (AUC) is the measure of a classifier’s ability to discriminate between classes and is used as a summary of the ROC curve. The higher the AUC value, the better the performance of the model in distinguishing between the positive and negative classes (Fawcett, 2004):

  $$AUC = \frac{1 + S_n - FP_{rate}}{2}$$

  where FP_{rate} is the ratio of the number of false positives to the total number of non-outlier days:

  $$FP_{rate} = \frac{FP}{TN + FP} = 1 - S_p$$

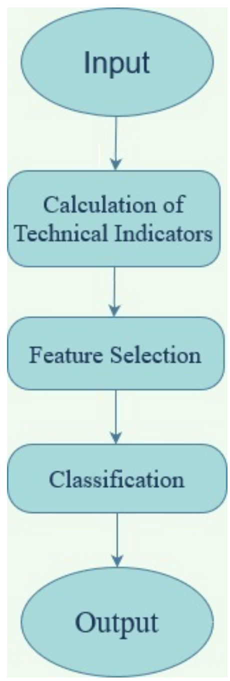
- The geometric mean (GM) of the sensitivity and specificity measures:

$$GM = \sqrt{S_n \times S_p}$$

- Matthew's correlation coefficient (MCC) provides better balance among the four basic metrics Sn, Sp, PPV and NPV (Matthews, 1975):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TP + FP) \times (TP + FN) \times (TN + FP)}}$$

Figure 1 Overall framework for the stock price trend prediction model (see online version for colours)



4 Experimental results

The results of the feature selection and classification algorithms applied to the data are presented in Tables 1 through 3 and Figures 2 to 19. In addition, Figure 20 illustrates the frequency of indicators selected by the feature selection algorithms in the daily period, and Figure 21 displays the frequency of indicators selected by the feature selection algorithms in the weekly period. Figures 20 and 21 reveal that the Senkou-span A, Senkou-span B, Kijun-Sen, MAOBV, and OBV indicators were most frequently selected by the feature selection algorithms, indicating that they may be particularly effective in detecting stock price trends. The implementations of our algorithms are available at http://faculty.du.ac.ir/rahmani/softwares/.

Table 1 Results of feature selection algorithms and SVM classification

Symbol – period	*Search method – eval model*	*Accuracy*	*MAE*	*RMSE*	*Selected features*	*AUC-ROC*	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*F-measure*	*GM*	*MCC*
Kpars – daily	CFS – BEST FIRST	56.735	0.433	0.658	6	0.568	0.551	0.585	0.592	0.544	0.571	0.568	0.136
Kpars – weekly	CFS – BEST FIRST	59.542	0.405	0.636	4	0.593	0.402	0.783	0.643	0.574	0.494	0.561	0.201
Valiz – daily	CFS – BEST FIRST	53.681	0.463	0.681	2	0.508	0.924	0.092	0.539	0.516	0.681	0.292	0.030
Valiz – weekly	CFS – BEST FIRST	54.853	0.452	0.672	7	0.548	0.568	0.529	0.550	0.546	0.559	0.548	0.097
Shbhran – daily	CFS – BEST FIRST	55.339	0.447	0.668	5	0.551	0.612	0.490	0.564	0.539	0.587	0.548	0.103
Shbhran – weekly	CFS – BEST FIRST	57.865	0.421	0.649	8	0.578	0.661	0.495	0.570	0.590	0.612	0.572	0.158
Kpars – daily	InfoGain – attribute ranking	57.484	0.425	0.652	25	0.576	0.547	0.605	0.603	0.550	0.573	0.575	0.152
Kpars – weekly	InfoGain – attribute ranking	59.542	0.405	0.636	14	0.595	0.547	0.642	0.598	0.594	0.571	0.593	0.190
Valiz – daily	InfoGain – attribute ranking	54.886	0.451	0.672	25	0.531	0.796	0.265	0.554	0.532	0.653	0.459	0.073
Valiz – weekly	InfoGain – attribute ranking	54.397	0.456	0.675	13	0.544	0.537	0.551	0.549	0.540	0.543	0.544	0.088
Shbhran – daily	InfoGain – attribute ranking	55.608	0.444	0.666	7	0.554	0.600	0.508	0.569	0.541	0.584	0.552	0.109
Shbhran – weekly	InfoGain – attribute ranking	57.292	0.427	0.654	15	0.573	0.617	0.528	0.570	0.576	0.593	0.571	0.146
Kpars – daily	GainRatio – attribute ranking	56.361	0.436	0.661	6	0.565	0.543	0.587	0.589	0.540	0.565	0.564	0.129
Kpars – weekly	GainRatio – attribute ranking	59.214	0.408	0.639	22	0.591	0.531	0.652	0.597	0.589	0.562	0.588	0.184
Valiz – daily	GainRatio – attribute ranking	54.886	0.451	0.672	25	0.531	0.796	0.265	0.554	0.532	0.653	0.459	0.073
Valiz – weekly	GainRatio – attribute ranking	55.049	0.450	0.671	15	0.551	0.548	0.553	0.555	0.546	0.551	0.551	0.101
Shbhran – daily	GainRatio – attribute ranking	55.642	0.444	0.666	7	0.555	0.600	0.509	0.569	0.541	0.584	0.553	0.110
Shbhran – weekly	GainRatio – attribute ranking	58.134	0.419	0.647	16	0.581	0.621	0.541	0.578	0.585	0.599	0.580	0.163
Kpars – daily	Correlation – attribute ranking	56.689	0.433	0.658	7	0.567	0.563	0.571	0.590	0.544	0.576	0.567	0.134
Kpars – weekly	Correlation – attribute ranking	58.887	0.411	0.641	10	0.588	0.541	0.635	0.590	0.588	0.565	0.586	0.177
Valiz – daily	Correlation – attribute ranking	54.821	0.452	0.672	18	0.530	0.802	0.257	0.553	0.531	0.655	0.454	0.071
Valiz – weekly	Correlation – attribute ranking	55.375	0.446	0.668	14	0.554	0.543	0.565	0.559	0.549	0.551	0.554	0.108
Shbhran – daily	Correlation – attribute ranking	55.406	0.446	0.668	9	0.552	0.598	0.506	0.567	0.539	0.582	0.550	0.105
Shbhran – weekly	Correlation – attribute ranking	56.821	0.432	0.657	12	0.568	0.621	0.515	0.565	0.572	0.591	0.565	0.136
Kpars – daily	Symmetrical – attribute ranking	56.735	0.433	0.658	7	0.568	0.555	0.581	0.592	0.544	0.573	0.568	0.136
Kpars – weekly	Symmetrical – attribute ranking	59.214	0.408	0.639	22	0.591	0.531	0.652	0.597	0.589	0.562	0.588	0.184
Valiz – daily	Symmetrical – attribute ranking	54.886	0.451	0.672	25	0.531	0.796	0.265	0.554	0.532	0.653	0.459	0.073
Valiz – weekly	Symmetrical – attribute ranking	55.179	0.448	0.670	15	0.552	0.548	0.555	0.556	0.548	0.552	0.552	0.104
Shbhran – daily	Symmetrical – attribute ranking	55.574	0.444	0.667	7	0.554	0.600	0.508	0.568	0.541	0.584	0.552	0.108
Shbhran – weekly	Symmetrical – attribute ranking	58.067	0.419	0.648	16	0.580	0.620	0.541	0.578	0.584	0.598	0.579	0.161
Kpars – daily	0	57.484	0.425	0.652	0	0.576	0.547	0.605	0.603	0.550	0.573	0.575	0.152
Kpars – weekly	0	60.851	0.392	0.626	0	0.608	0.552	0.664	0.614	0.604	0.581	0.605	0.217
Valiz – daily	0	54.886	0.451	0.672	0	0.531	0.796	0.265	0.554	0.532	0.653	0.459	0.073
Valiz – weekly	0	55.342	0.447	0.668	0	0.553	0.548	0.559	0.558	0.549	0.553	0.553	0.107
Shbhran – daily	0	56.315	0.437	0.661	0	0.562	0.581	0.544	0.579	0.546	0.580	0.562	0.125
Shbhran – weekly	0	58.875	0.411	0.641	0	0.589	0.623	0.554	0.586	0.592	0.604	0.587	0.177

Table 2 Results of feature selection algorithms and C4.5 classification

Symbol – period	*Search method – eval model*	*Accuracy*	*MAE*	*RMSE*	*Selected features*	*AUC-ROC*	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*F-measure*	*GM*	*MCC*
Kpars – daily	CFS – BEST FIRST	57.998	0.454	0.487	6	0.620	0.379	0.800	0.675	0.541	0.485	0.550	0.196
Kpars – weekly	CFS – BEST FIRST	69.738	0.356	0.455	4	0.759	0.614	0.778	0.729	0.675	0.667	0.691	0.398
Valiz – daily	CFS – BEST FIRST	55.114	0.477	0.490	2	0.566	0.963	0.078	0.545	0.651	0.696	0.275	0.091
Valiz – weekly	CFS – BEST FIRST	70.782	0.338	0.455	7	0.767	0.729	0.687	0.703	0.714	0.715	0.707	0.416
Shbhran – daily	CFS – BEST FIRST	54.261	0.487	0.496	5	0.564	0.426	0.668	0.581	0.519	0.492	0.534	0.097
Shbhran – weekly	CFS – BEST FIRST	72.449	0.323	0.455	8	0.773	0.729	0.720	0.725	0.724	0.727	0.724	0.449
Kpars – daily	InfoGain – attribute ranking	58.092	0.440	0.491	25	0.635	0.690	0.461	0.584	0.576	0.632	0.564	0.156
Kpars – weekly	InfoGain – attribute ranking	69.504	0.341	0.464	14	0.754	0.583	0.804	0.742	0.665	0.653	0.685	0.397
Valiz – daily	InfoGain – attribute ranking	56.091	0.463	0.508	25	0.589	0.524	0.603	0.602	0.525	0.561	0.562	0.127
Valiz – weekly	InfoGain – attribute ranking	72.671	0.303	0.454	13	0.784	0.745	0.708	0.722	0.732	0.733	0.726	0.453
Shbhran – daily	InfoGain – attribute ranking	53.857	0.487	0.497	7	0.558	0.487	0.595	0.564	0.518	0.523	0.538	0.082
Shbhran – weekly	InfoGain – attribute ranking	73.190	0.299	0.455	15	0.785	0.716	0.748	0.742	0.722	0.729	0.732	0.464
Kpars – daily	GainRatio – attribute ranking	55.847	0.463	0.489	6	0.595	0.378	0.756	0.629	0.526	0.472	0.534	0.144
Kpars – weekly	GainRatio – attribute ranking	69.785	0.323	0.479	22	0.753	0.625	0.769	0.724	0.679	0.671	0.693	0.398
Valiz – daily	GainRatio – attribute ranking	56.091	0.463	0.508	25	0.589	0.524	0.603	0.602	0.525	0.561	0.562	0.127
Valiz – weekly	GainRatio – attribute ranking	72.638	0.311	0.459	15	0.775	0.747	0.706	0.721	0.733	0.733	0.726	0.453
Shbhran – daily	GainRatio – attribute ranking	53.857	0.487	0.497	7	0.558	0.487	0.595	0.564	0.518	0.523	0.538	0.082
Shbhran – weekly	GainRatio – attribute ranking	73.931	0.289	0.456	16	0.788	0.733	0.746	0.745	0.734	0.739	0.739	0.479
Kpars – daily	GainRatio – attribute ranking	56.408	0.461	0.490	7	0.597	0.326	0.825	0.670	0.528	0.439	0.518	0.173
Kpars – weekly	GainRatio – attribute ranking	64.921	0.391	0.478	10	0.698	0.478	0.816	0.716	0.617	0.573	0.624	0.312
Valiz – daily	Correlation – attribute ranking	55.375	0.470	0.508	18	0.581	0.545	0.564	0.589	0.519	0.566	0.554	0.109
Valiz – weekly	Correlation – attribute ranking	66.189	0.366	0.477	14	0.726	0.613	0.711	0.683	0.644	0.646	0.660	0.326
Shbhran – daily	Correlation – attribute ranking	54.261	0.481	0.494	9	0.566	0.391	0.707	0.590	0.518	0.470	0.525	0.102
Shbhran – weekly	Correlation – attribute ranking	69.754	0.332	0.475	12	0.751	0.665	0.731	0.715	0.683	0.689	0.697	0.396
Kpars – daily	Symmetrical – attribute ranking	57.624	0.457	0.489	7	0.614	0.453	0.711	0.632	0.543	0.528	0.568	0.169
Kpars – weekly	Symmetrical – attribute ranking	69.832	0.322	0.478	22	0.753	0.625	0.770	0.725	0.679	0.671	0.693	0.399
Valiz – daily	Symmetrical – attribute ranking	56.091	0.463	0.508	25	0.589	0.524	0.603	0.602	0.525	0.561	0.562	0.127
Valiz – weekly	Symmetrical – attribute ranking	72.606	0.311	0.459	15	0.775	0.747	0.705	0.720	0.733	0.733	0.726	0.452
Shbhran – daily	Symmetrical – attribute ranking	53.857	0.487	0.497	7	0.558	0.487	0.595	0.564	0.518	0.523	0.538	0.082
Shbhran – weekly	Symmetrical – attribute ranking	73.998	0.287	0.456	16	0.789	0.733	0.747	0.746	0.734	0.740	0.740	0.480
Kpars – daily	0	58.092	0.440	0.491	0	0.635	0.690	0.461	0.584	0.576	0.632	0.564	0.156
Kpars – weekly	0	74.135	0.279	0.471	0	0.769	0.727	0.755	0.742	0.741	0.735	0.741	0.483
Valiz – daily	0	56.091	0.463	0.508	0	0.589	0.524	0.603	0.602	0.525	0.561	0.562	0.127
Valiz – weekly	0	72.932	0.300	0.474	0	0.758	0.754	0.704	0.721	0.738	0.737	0.729	0.459
Shbhran – daily	0	54.362	0.477	0.495	0	0.572	0.528	0.560	0.564	0.524	0.546	0.544	0.088
Shbhran – weekly	0	74.503	0.282	0.464	0	0.778	0.748	0.742	0.746	0.744	0.747	0.745	0.490

Table 3 Results of feature selection algorithms and random forest classification

Symbol – period	*Search method – eval model*	*Accuracy*	*MAE*	*RMSE*	*Selected features*	*AUC-ROC*	*Sensitivity*	*Specificity*	*PPV*	*NPV*	*F-measure*	*GM*	*MCC*
Kpars – daily	CFS – BEST FIRST	61.459	0.418	0.478	6	0.682	0.636	0.592	0.630	0.597	0.633	0.613	0.227
Kpars – weekly	CFS – BEST FIRST	83.302	0.259	0.353	4	0.902	0.823	0.842	0.835	0.831	0.829	0.833	0.666
Valiz – daily	CFS – BEST FIRST	53.420	0.469	0.548	2	0.560	0.575	0.487	0.563	0.500	0.569	0.529	0.063
Valiz – weekly	CFS – BEST FIRST	82.541	0.276	0.360	7	0.897	0.818	0.833	0.832	0.819	0.825	0.825	0.651
Shbhran – daily	CFS – BEST FIRST	57.359	0.463	0.500	5	0.602	0.618	0.525	0.584	0.561	0.601	0.570	0.144
Shbhran – weekly	CFS – BEST FIRST	85.416	0.253	0.333	8	0.926	0.870	0.839	0.845	0.864	0.857	0.854	0.709
Kpars – daily	InfoGain – attribute ranking	63.751	0.412	0.468	25	0.703	0.667	0.605	0.649	0.624	0.658	0.635	0.273
Kpars – weekly	InfoGain – attribute ranking	84.425	0.266	0.343	14	0.917	0.841	0.847	0.842	0.846	0.842	0.844	0.688
Valiz – daily	InfoGain – attribute ranking	60.326	0.440	0.482	25	0.654	0.649	0.550	0.624	0.578	0.636	0.598	0.201
Valiz – weekly	InfoGain – attribute ranking	82.117	0.289	0.361	13	0.901	0.830	0.812	0.818	0.825	0.824	0.821	0.642
Shbhran – daily	InfoGain – attribute ranking	57.460	0.463	0.497	7	0.602	0.635	0.509	0.583	0.564	0.608	0.569	0.146
Shbhran – weekly	InfoGain – attribute ranking	84.170	0.280	0.345	15	0.921	0.852	0.831	0.837	0.847	0.844	0.842	0.683
Kpars – daily	GainRatio – attribute ranking	62.722	0.421	0.479	6	0.677	0.654	0.598	0.640	0.612	0.647	0.625	0.252
Kpars – weekly	GainRatio – attribute ranking	84.425	0.283	0.351	22	0.913	0.833	0.855	0.848	0.841	0.840	0.844	0.688
Valiz – daily	GainRatio – attribute ranking	60.326	0.440	0.482	25	0.654	0.649	0.550	0.624	0.578	0.636	0.598	0.201
Valiz – weekly	GainRatio – attribute ranking	80.977	0.300	0.367	15	0.896	0.814	0.805	0.809	0.810	0.812	0.810	0.620
Shbhran – daily	GainRatio – attribute ranking	56.787	0.465	0.498	7	0.600	0.616	0.515	0.579	0.555	0.597	0.564	0.133
Shbhran – weekly	GainRatio – attribute ranking	85.787	0.271	0.339	16	0.926	0.869	0.847	0.852	0.864	0.860	0.858	0.716
Kpars – daily	Correlation – attribute ranking	59.355	0.437	0.483	7	0.652	0.624	0.560	0.608	0.577	0.616	0.591	0.185
Kpars – weekly	Correlation – attribute ranking	79.607	0.330	0.386	10	0.878	0.786	0.806	0.797	0.795	0.792	0.796	0.592
Valiz – daily	Correlation – attribute ranking	60.033	0.443	0.482	18	0.651	0.647	0.547	0.621	0.575	0.634	0.595	0.195
Valiz – weekly	Correlation – attribute ranking	79.674	0.323	0.381	14	0.881	0.796	0.798	0.800	0.794	0.798	0.797	0.593
Shbhran – daily	Correlation – attribute ranking	57.730	0.460	0.492	9	0.614	0.643	0.506	0.584	0.568	0.612	0.571	0.151
Shbhran – weekly	Correlation – attribute ranking	83.766	0.296	0.357	12	0.911	0.852	0.824	0.830	0.845	0.841	0.837	0.675
Kpars – daily	Symmetrical – attribute ranking	63.003	0.413	0.476	7	0.692	0.651	0.607	0.645	0.614	0.648	0.629	0.258
Kpars – weekly	Symmetrical – attribute ranking	83.489	0.284	0.353	22	0.911	0.824	0.845	0.838	0.832	0.831	0.835	0.670
Valiz – daily	Symmetrical – attribute ranking	60.326	0.440	0.482	25	0.654	0.649	0.550	0.624	0.578	0.636	0.598	0.201
Valiz – weekly	Symmetrical – attribute ranking	81.661	0.299	0.367	15	0.896	0.825	0.808	0.813	0.820	0.819	0.816	0.633
Shbhran – daily	Symmetrical – attribute ranking	56.046	0.464	0.498	7	0.599	0.617	0.499	0.571	0.547	0.593	0.555	0.117
Shbhran – weekly	Symmetrical – attribute	85.719	0.273	0.339	16	0.926	0.868	0.847	0.852	0.863	0.860	0.857	0.714
Kpars – daily	No filter	63.751	0.412	0.468	0	0.703	0.667	0.605	0.649	0.624	0.658	0.635	0.273
Kpars – weekly	No filter	83.536	0.281	0.351	0	0.912	0.822	0.848	0.840	0.831	0.831	0.835	0.671
Valiz – daily	No filter	60.326	0.440	0.482	0	0.654	0.649	0.550	0.624	0.578	0.636	0.598	0.201
Valiz – weekly	No filter	81.857	0.308	0.369	0	0.897	0.829	0.808	0.814	0.823	0.822	0.818	0.637
Shbhran – daily	No filter	60.593	0.439	0.481	0	0.660	0.640	0.569	0.616	0.594	0.628	0.604	0.210
Shbhran – weekly	No filter	85.349	0.281	0.344	0	0.923	0.870	0.836	0.844	0.864	0.857	0.853	0.707

Figure 2 Results of SVM classifier without the use of feature selection algorithms (see online version for colours)

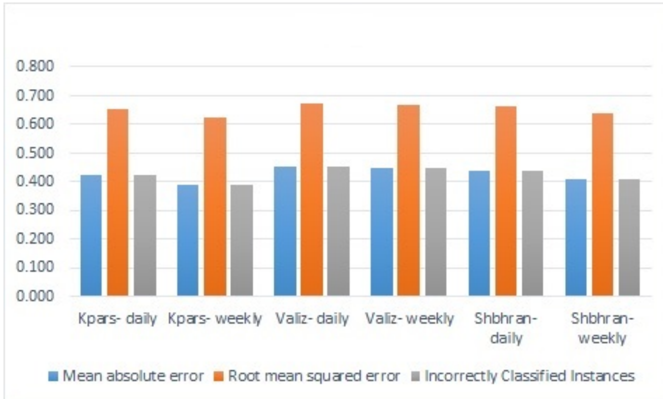


Figure 3 Results of using the symmetrical attribute evaluator model with an SVM classifier (see online version for colours)

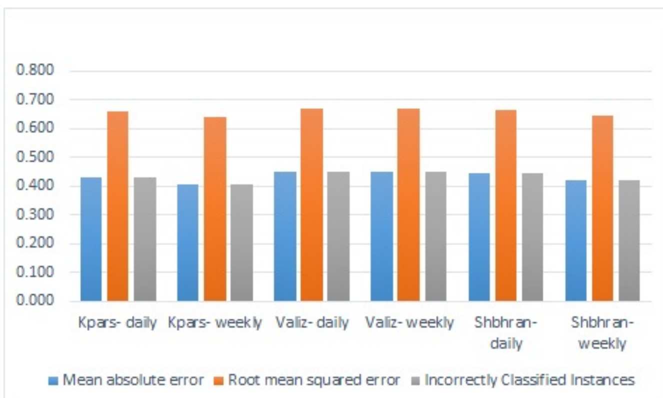


Figure 4 Results of using the correlation attribute evaluator model with an SVM classifier (see online version for colours)

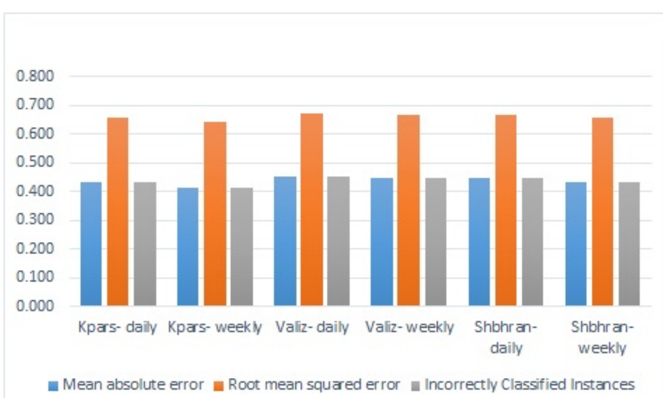


Figure 5 Results of using the gain ratio feature evaluator model with an SVM classifier (see online version for colours)

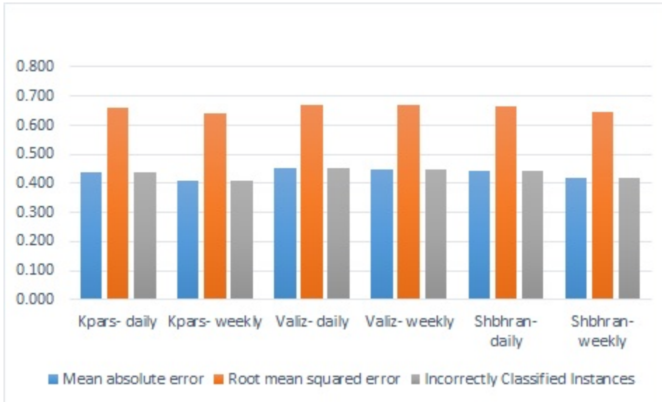


Figure 6 Results of using the information gain attribute evaluator model with an SVM classifier (see online version for colours)

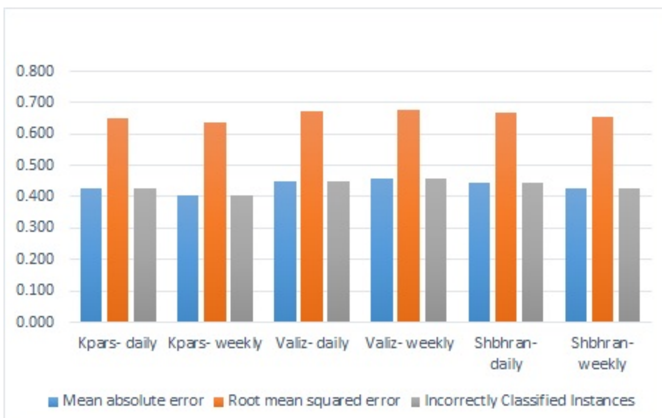


Figure 7 Results of using the CFS subset evaluator model with an SVM classifier (see online version for colours)

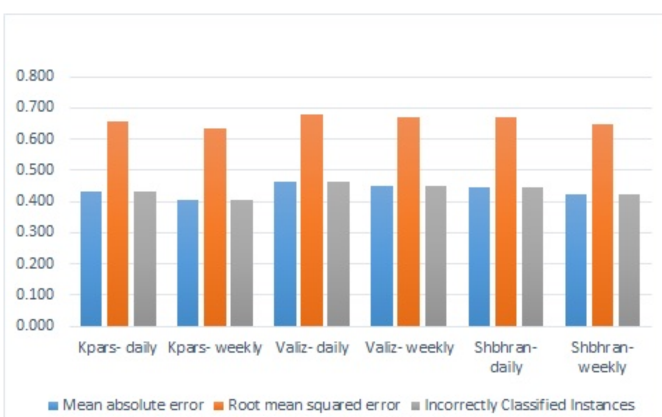


Figure 8 Results of C4.5 classifier without the use of feature selection algorithms (see online version for colours)

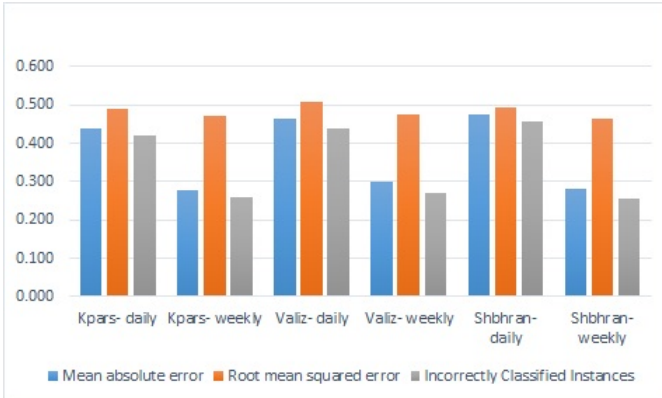


Figure 9 Results of using the symmetrical attribute evaluator model with a C4.5 classifier (see online version for colours)

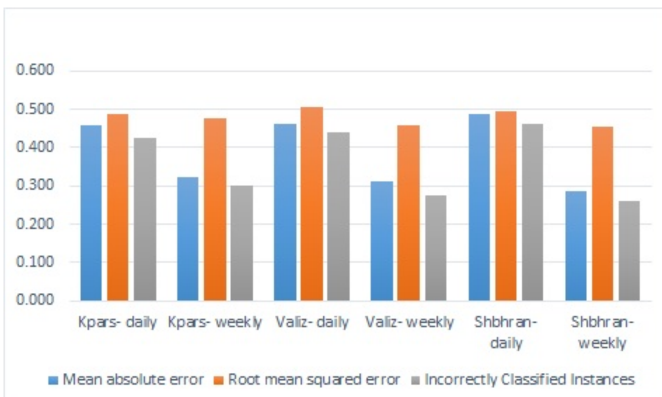


Figure 10 Results of using the correlation attribute evaluator model with a C4.5 classifier (see online version for colours)

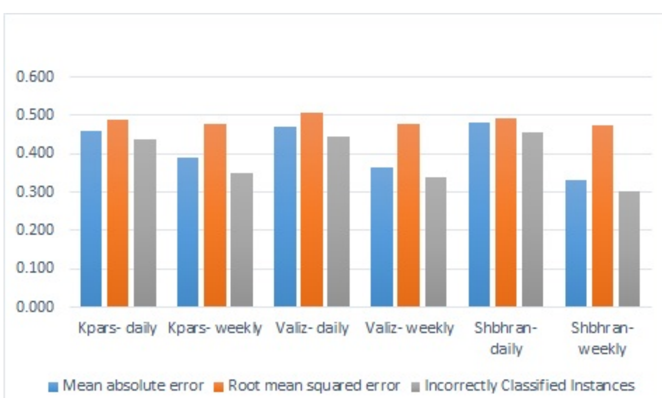


Figure 11 Results of using the gain ratio feature evaluator model with a C4.5 classifier (see online version for colours)

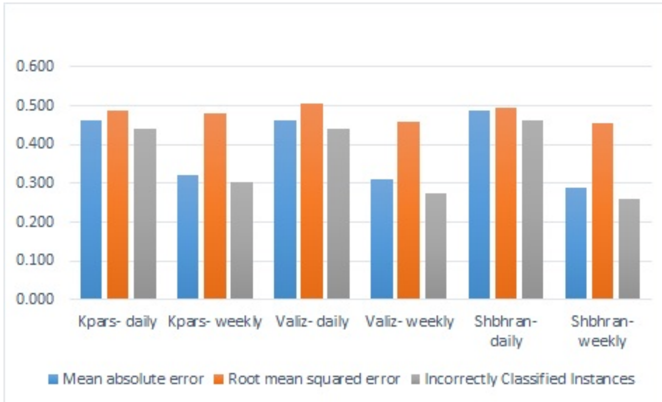


Figure 12 Results of using the information gain attribute evaluator model with a C4.5 classifier (see online version for colours)

Figure 13 Results of using the CFS subset evaluator model with a C4.5 classifier (see online version for colours)

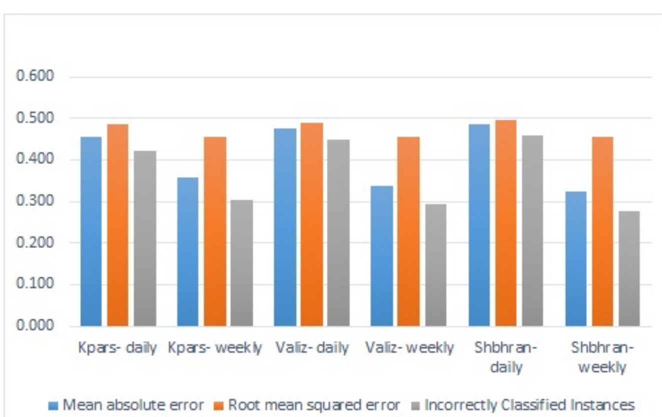


Figure 14 Results of random forest classifier without the use of feature selection algorithms (see online version for colours)

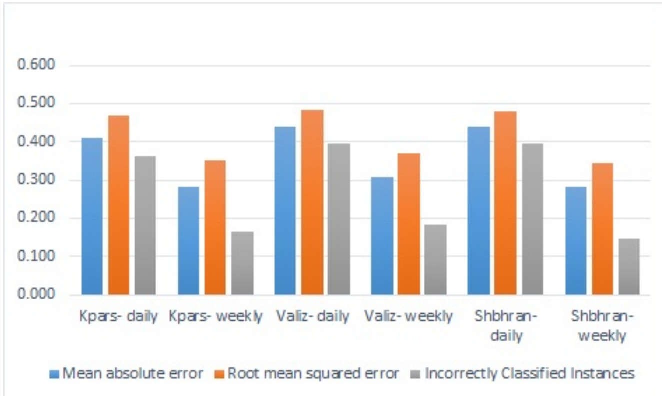


Figure 15 Results of using the symmetrical attribute evaluator model with a random forest classifier (see online version for colours)

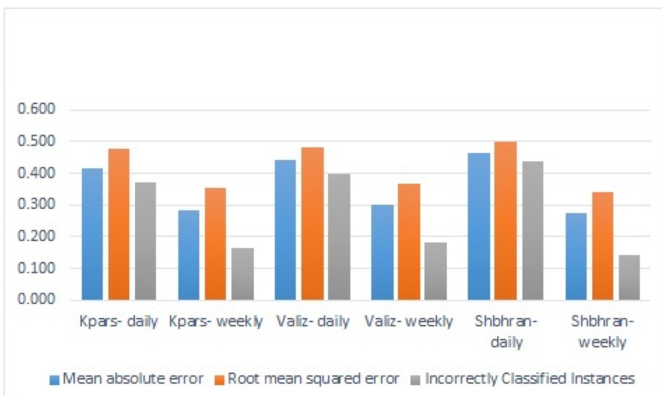


Figure 16 Results of using the correlation attribute evaluator model with a random forest classifier (see online version for colours)

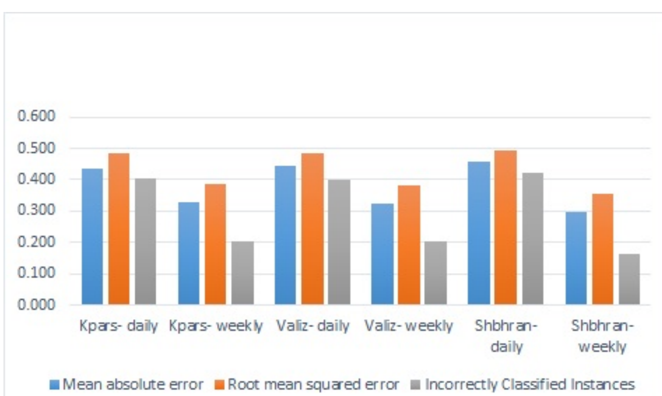


Figure 17 Results of using the gain ratio feature evaluator model with a random forest classifier (see online version for colours)

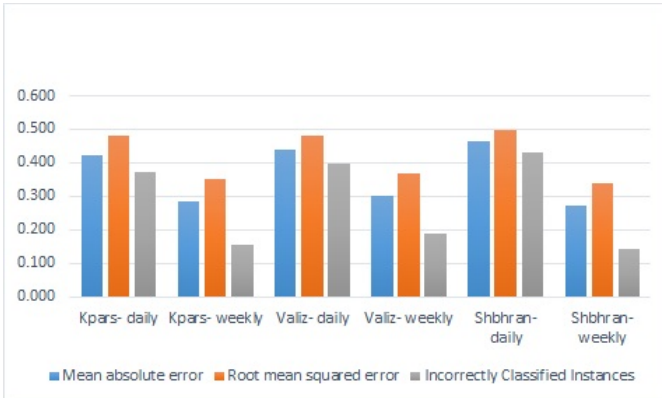


Figure 18 Results of using the information gain attribute evaluator model with a random forest classifier (see online version for colours)

Figure 19 Results of using the CFS subset evaluator model with a random forest classifier (see online version for colours)

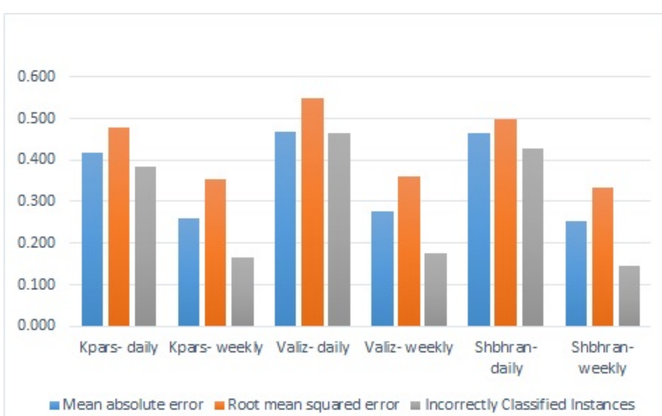


Figure 20 Frequency of the indicators selected by feature selection algorithms in a daily period (see online version for colours)

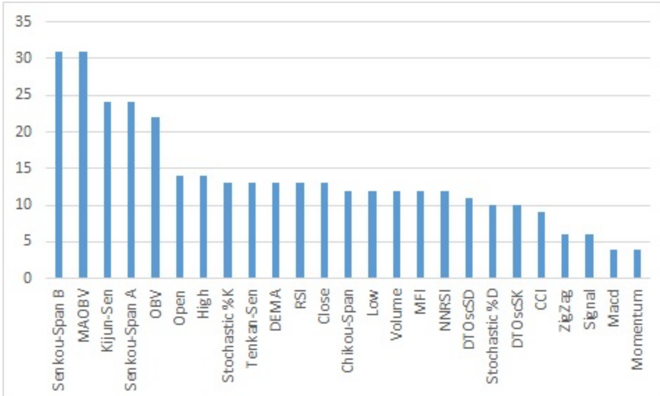
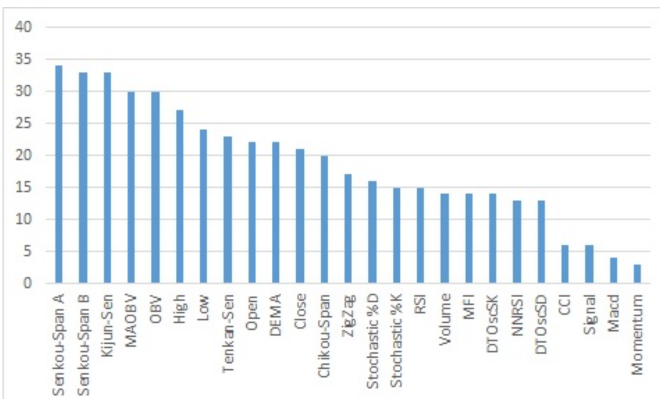


Figure 21 Frequency of the indicators selected by feature selection algorithms in a weekly period (see online version for colours)



5 Discussion

Based on the experimental results, it has been observed that future trading trends can be predicted with a high degree of accuracy using less than 25 features. In some cases, it was found that better accuracy was achieved when using fewer features compared to using all available features. For instance, when predicting the 'Shbhran' trend, the GainRatio-attribute ranking feature selection algorithm in combination with random forest classification achieved an accuracy of 85.787% using only 16 selected features (close, MAOBV, OBV, Chikou-span, RSI, signal, DEMA, Macd, Senkou-span B, NNRSI, MFI, Senkou-span A, CCI, stochastic %K, momentum, stochastic %D) in the weekly period, whereas the accuracy was slightly lower at 85.349% when using all 25 features without applying a feature selection algorithm. Similarly, when predicting the 'Kpars' trend, the InfoGain-attribute ranking feature selection algorithm in combination with random forest classification achieved an accuracy of 84.425% using only 14 selected features (MAOBV, Kijun-Sen, OBV, Senkou-span A, Senkou-span B, DEMA, Tenkan-Sen, RSI, MFI, low, stochastic %K, CCI, stochastic %D, zigzag) in the weekly period, whereas the accuracy was slightly lower at 83.536% when using all 25 features without applying a feature selection algorithm. These results demonstrate the effectiveness of the proposed model in accurately predicting stock price trends using a reduced set of features.

6 Conclusions

In this study, we investigated the effectiveness of feature selection and classification techniques in forecasting stock price changes on the Tehran Stock Exchange. We examined 25 different features and indicators and made predictions on a daily and weekly basis. Our experimental findings revealed that the random forest classification algorithm outperformed other methods in several instances, demonstrating superior accuracy. Furthermore, by reducing the number of features and indicators, we achieved even higher prediction accuracy, particularly for weekly forecasts, while simultaneously reducing computational burden. Notably, our results indicated that the accuracy of weekly price trend predictions surpassed that of daily price trend predictions.

The findings of this study highlight the potential benefits of employing feature selection and classification methods in stock price trend prediction. These techniques can assist investors in achieving more precise forecasts, leading to a reduction in investment risk and an increase in profits. By carefully selecting a subset of pertinent and informative features, it becomes possible to enhance the accuracy of predictions while simplifying the computational process. This approach holds particular value for investors seeking to make well-informed decisions regarding their stock portfolio.

Based on the results of this work, it is suggested that future research could explore the use of additional feature selection and classification algorithms, or apply the algorithms tested in this research to data from other markets such as the gold or cryptocurrency markets. This would provide a more comprehensive understanding of the effectiveness of these algorithms in different contexts and could help identify potential improvements or modifications to the existing approach. Additionally, it may be valuable to investigate the use of more advanced machine learning techniques, such as deep learning, to further improve the accuracy and efficiency of stock trend prediction models. Overall, there is significant potential for continued research in this area to yield valuable insights and improve the performance of stock market analysis and prediction.

References

Appel, G. (2005) *Technical Analysis: Power Tools for Active Investors*, FT Press.

Biau, G. and Scornet, E. (2016) 'A random forest guided tour', *Test*, Vol. 25, No. 2, pp.197–227.

Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. (2016) *WEKA Manual for Version 3-8-1*, University of Waikato, Hamilton, New Zealand.

Bramer, M. (2007) *Principle of Data Mining*, Springer.

Chandar, S.K. (2020) 'Grey wolf optimization-Elman neural network model for stock price prediction', *Soft Computing*, Vol. 25, No. 1, pp.649–658.

Chen, L., Qiao, Z., Wang, M., Wang, C., Du, R. and Stanley, H.E. (2018) 'Which artificial intelligence algorithm better predicts the Chinese stock market?', *IEEE Access*, Vol. 6, No. 1, pp.48625–48633.

Chong, T.T.L. and Ng, W.K. (2008) 'Technical analysis and the London Stock Exchange: testing the MACD and RSI rules using the FT30', *Applied Economics Letters*, Vol. 15, No. 14, pp.1111–1114.

Chung, H. and Shin, K.S. (2018) 'Genetic algorithm-optimized long short-term memory network for stock market prediction', *Sustainability*, Vol. 10, No. 10, pp.1–18.

Dash, M. and Liu, H. (1997) 'Feature selection for classification', *Intelligent Data Analysis*, Vol. 1, pp.131–156.

Dash, R. and Dash, P.K. (2016) 'A hybrid stock trading framework integrating technical analysis with machine learning techniques', *The Journal of Finance and Data Science*, Vol. 2, No. 1, pp.42–57.

Diaz, F. and Jones, R. (2004) 'Using temporal profiles of queries for precision prediction', *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.18–24.

Di Lorenzo, R. (2012) *How to Make Money by Fast Trading: A Guide to Success*, Springer Science & Business Media, New York.

Fawcett, T. (2004) 'ROC graphs: notes and practical considerations for researchers', *Machine Learning*, Vol. 31, No. 1, pp.1–38.

Fayek, M.B., El-Boghdadi, H.M. and Omran, S.M. (2013) 'Multi-objective optimization of technical stock market indicators using gas', *International Journal of Computer Applications*, Vol. 68, No. 20.

Guo, Z.Q., Wang, H.Q. and Liu, Q. (2013) 'Financial time series forecasting using LPP and SVM optimized by PSO', *Soft Comput.*, Vol. 7, No. 5, pp.805–818.

Ian, H.W. and Eibe, F. (2005) *Data Mining Practical Machine Learning Tools and Techniques*.

Kantardzic, M. (2003) *Data Mining: Concepts, Models, Methods, and Algorithms*, Wiley-Interscience.

Kara, Y., Boyacioglu, M.A. and Baykan, O.K. (2011) 'Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul Stock Exchange', *Expert Syst. Appl.*, Vol. 38, No. 5, pp.5311–5319.

Lei, L. (2018) 'Wavelet neural network prediction method of stock price trend based on rough set attribute reduction', *Appl. Soft Comput.*, Vol. 62, No. 1, pp.923–932.

Liao, Z. and Wang, J. (2010) 'Forecasting model of global stock index by stochastic time effective neural networks', *Expert Syst. Appl.*, Vol. 37, No. 1, pp.834–841.

Lin, X., Yang, Z. and Song, Y. (2009) 'Short-term stock price prediction based on echo state networks', *Expert Syst. Appl.*, Vol. 36, No. 3, pp.7313–7317.

Marek, P. and Cadkova, V. (2020) 'Optimization and testing of money flow index', *APLIMAT 2020 – 19th Conference on Applied Mathematics*.

Matthews, B.W. (1975) 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme', *Biochimica et Biophysica Acta (BBA) – Protein Structure*, Vol. 405, No. 2, pp.442–451.

Miner, R.C. (2008) *High Probability Trading Strategies: Entry to Exit Tactics for the Forex, Futures, and Stock Markets*, Vol. 328, John Wiley & Sons.

Murphy, J.J. (1999) *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, New York Institute of Finance Series, New York Institute of Finance.

Murphy, J.J. (2009) *The Visual Investor: How to Spot Market Trends*, 2nd ed., John Wiley & Sons, Hoboken, New Jersey.

Nahil, A. and Lyhyaoui, A. (2018) 'Short-term stock price forecasting using kernel principal component analysis and support vector machines: the case of Casablanca Stock Exchange', *Procedia Comput. Sci.*, Vol. 127, No. 1, pp.161–169.

Ntakaris, A., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2020) 'Mid-price prediction based on machine learning methods with technical and quantitative indicators', *Plos One*, Vol. 15, No. 6, p.e0234107.

Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W. and Wills, G. (2020) 'A survey on machine learning for stock price prediction: algorithms and techniques', in *Proceedings of the 2nd International Conference on Finance, Economics, Management and IT Business – FEMIB 2020*, Prague, Czech Republic, pp.63–71.

Omuya, E.O., Okeyo, G.O. and Kimwele, M.W. (2021) 'Feature selection for classification using principal component analysis and information gain', *Expert Systems with Applications*, Vol. 174, No. 1, p.114765.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015) 'Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques', *Expert Systems with Applications*, Vol. 42, No. 1, pp.259–268.

Patel, M. (2010) *Trading with Ichimoku Clouds: The Essential Guide to Ichimoku Kinko Hyo Technical Analysis*, Vol. 473, John Wiley & Sons, Hoboken, New Jersey.

Quinlan, J.R. (2014) *C4.5: Programs for Machine Learning*, pp.58–60, Elsevier.

Rahimunnisa, K. (2019) 'Hybrdized genetic-simulated annealing algorithm for performance optimization in wireless adhoc network', *J. Soft Comput Paradigm (JSCP)*, Vol. 1, No. 1, pp.1–13.

Ray, P., Mahapatra, G.S., Rani, P., Pandy, S.K. and Dey, K.N. (2014) 'Robust feed forward and recurrent neural network based dynamic weighted combination models for software reliability prediction', *Appl. Soft. Comput.*, Vol. 22, No. 1, pp.629–637.

Shi, H. and Liu, X. (2014) 'Application on stock price prediction on Elman neural networks based on principal component method', *2014 International Computer Conference on Wavelet Active Technology and Information Processing*.

Swales, G. and Yoon, Y. (1991) 'Predicting stock price performance: a neural network approach', in *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, IEEE, Vol. 4, No. 1, pp.156–162.

Talebi, H., Hoang, W. and Gavrilova, M.L. (2014) 'Multi-scale foreign exchange rates ensemble for classification of trends in forex market', *Procedia Computer Science*, Vol. 29, No. 1, pp.2065–2075.

Tsang, W.W.H. and Chong, T.T.L. (2009) 'Profitability of the on-balance volume indicator', *Economics Bulletin*, Vol. 29, No. 3, pp.2424–2431.

Vandana, C.P. and Chikkamannur, A.A. (2021) 'Feature selection: an empirical study', *International Journal of Engineering Trends and Technology*, Vol. 69, No. 2, pp.165–170.

Witten, I.H., Frank, E. and Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann Publishers, Burlington.

Wong, H.B. and Lim, G.H. (2011) 'Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV', *Proceedings of Singapore Healthcare*, Vol. 20, No. 4, pp.316–318.

Xu, Y., Chhim, L., Zheng, B. and Nojima, Y. (2020) 'Stacked deep learning structure with bidirectional long-short term memory for stock market prediction', *International Conference on Neural Computing for Advanced Applications*, July, Springer, Singapore, pp.447–460.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A. and Liu, H. (2010) 'Advancing feature selection research', *ASU Feature Selection Repository*, pp.1–28.

Zhong, X. and Enke, D. (2017) 'Forecasting daily stock market return using dimensionality reduction', *Expert Syst. Appl.*, Vol. 67, No. 1, pp.126–139.