data

MDPI

Article

Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining

Manik Sharma [1,*], Samriti Sharma [2] and Gurvinder Singh [2]

[1] Department of Computer Science and Applications, DAV University, Jalandhar 144401, India
[2] Department of Computer Science, Guru Nanak Dev University, Amritar 143001, India; smritigndu@gmail.com (S.S.); gurvinder.dcse@gndu.ac.in (G.S.)
* Correspondence: manik_sharma25@yahoo.com

Received: 30 September 2018; Accepted: 19 November 2018; Published: 24 November 2018

Abstract: Nowadays, overwhelming stock data is available, which areonly of use if it is properly examined and mined. In this paper, the last twelve years of ICICI Bank's stock data have been extensively examined using statistical and supervised learning techniques. This study may be of great interest for those who wish to mine or study the stock data of banks or any financial organization. Different statistical measures have been computed to explore the nature, range, distribution, and deviation of data. The different descriptive statistical measures assist in finding different valuable metrics such as mean, variance, skewness, kurtosis, *p*-value, a-squared, and 95% confidence mean interval level of ICICI Bank's stock data. Moreover, daily percentage changes occurring over the last 12 years have also been recorded and examined. Additionally, the intraday stock status has been mined using ten different classifiers. The performance of different classifiers has been evaluated on the basis of various parameters such as accuracy, misclassification rate, precision, recall, specificity, and sensitivity. Based upon different parameters, the predictive results obtained using logistic regression are more acceptable than the outcomes of other classifiers, whereas naïve Bayes, C4.5, random forest, linear discriminant, and cubic support vector machine (SVM) merely act as a random guessing machine. The outstanding performance of logistic regression has been validated using TOPSIS (technique for order preference by similarity to ideal solution) and WSA (weighted sum approach).

Keywords: stock forecasting; naïve Bayes; C4.5; random forest; logistic regression; support vector machine

1. Introduction

The deep statistical analytics of a bank's stock data, along with the performance analysis of different classifiers, can significantly assist a financial analyst and data scientist in predicting intraday, weekly, monthly, and future values of the stock. In this manuscript, the stock data of ICICI bank has been examined using several statistical and supervised learning techniques. ICICI Bank is one of the important leading Indian private banks, comprised of more than 4000 branches that operate in 19 different countries [1]. The bank offers a range of financial services related to savings, current, and fixed deposit accounts. It also offers a different range of loans to rural and urban customers. In the last few years, ICICI bank has gained a lot of faith and confidence from its customers. This statement can be verified from the TRA brand trust report 2018, which declared ICICI to be top of the private Indian banks [2]. Figure 1 represents some of the key figures of the ICICI Banks. It is observed that when compared with the financial year (FY) of 2016, the year of 2017 has a good rate of net interest margins. Growth of 0.71% has been observed in 2017. For the same period, net NPA's stood at 4.89% (2.67% for FY 2016), which is slightly higher than the industry average for private banks.

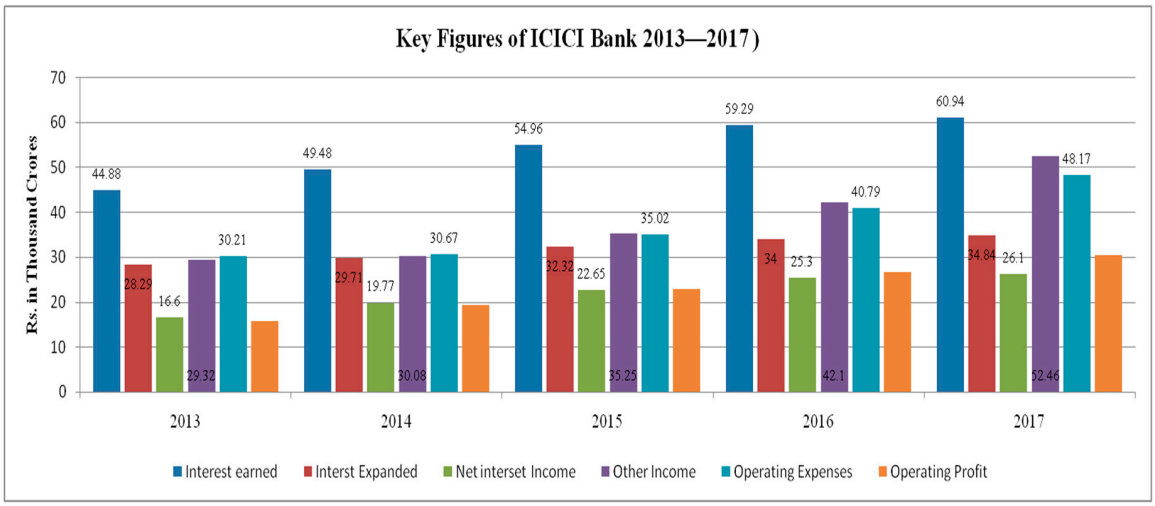

Figure 1. Key figures of ICICI Bank's stock. FY—financial year.

Nowadays, the momentous volume of data has been produced daily by different commercial, administrative, and scientific organizations that are tremendously expanding the sizes of the databases. Devices based on social media, the Internet, and the Internet of Things (IoT) further assist in fueling the growth of these databases [3]. Therefore, data are not an issue now. However, the real challenge lies in transforming this mountain of data into useful information. The use of statistics and data mining techniques assists in the automatic extraction of veiled data patterns that are otherwise buried in unprocessed data [4]. Data mining is a systematic data processing approach that collects, cleans, processes, examines, and extracts substantial qualities and hidden patterns in data [5,6]. In other words, the use of data mining techniques assists in extracting precise and significant insights from the deluge of data that have been collected from several manual and digital data sources [4,7]. Data mining has been used in several domains such as agriculture [7–9], finance [4,10–12], medical science [13–17], and bio-informatics [18,19].

Chandralekha and Shenbagavadivu have examined the performance of different machine learning techniques used to predict cardiovascular disorders. The study has been carried to compare and contrast the performances of both supervised and unsupervised learning techniques based upon three different metrics, namely, accuracy, recall, and precision. Authors found that the results obtained using decision trees are more accurate than those using other machine learning techniques [20]. Belavagi and B. Muniyal have evaluated the performance of different machine learning techniques employed for intrusion detection. Authors found that the results obtained in identifying intrusion using random forest outperform other those using machine learning techniques [21].

Previously, different authors have tried to examine the financial and multivariate analysis of ICICI Bank [22,23]. However, no significant work has been carried out to rigorously analyze ICICI Bank's stock using a combination of descriptive statistics and supervised learning techniques. The aggregated motive of this study is to extensively mine and analyze the data of one of the leading private Indian banks (ICICI, Mumbai, India) using descriptive and supervised learning techniques. Statistical techniques have been employed to examine the nature, trend, variation, and distribution of data. The last five years' key figures of ICICI Bank's stock have been examined to assess the economic status of ICICI Bank. Different descriptive statistical measuressuch as mean, standard deviation, variance, skewness, and kurtosis, along with *p*-values and A-squared values, have been computed for the major attributes of ICICI Bank's stock data. The last twelve years' daily deviation in the opening value of the stock has been recorded and analyzed to examine the intraday status of ICICI Bank's stock. Moreover, ten different classifiers, namely, naïve Bayes; C4.5; random forest; logistic regression; linear discriminant; and linear, quadratic, cubic, fine, and medium Gaussian support vector machines, have been used to classify the intraday status of ICICI Bank's stock data. The performances of the

different classifiers have been computed using eight different performance metrics. Additionally, the rate of misclassification, as well as of sensitivity along the F1-score, has also been computed and examined.

This study will be beneficial for financial analysts and researchers who wish to extensively mine the financial data of different leading Indian private or government banks. Moreover, it can be of potential interest to quantitative traders.

2. Related Work

Numerous researchers have strived to classify and forecast the future value of stocks using different statistical, data mining, and soft computing techniques. The stock and stock index has been predicted by trend deterministic data and machine learning by J. Patel [24]. Authors compared and contrasted the performance of four different machine learning approaches, that is, SVM (support vector machine), RF (random forest), NB (naïve Bayes), and ANN (artificial neural network), in predicting the future value for Reliance and Infosys, and found that the predictive rate is increased if the different trading parameters are represented as deterministic trend data. Al-Radqidah et al. have predicted the stock price of three different enterprises of the Amman Stock Exchange using ID3 and C4.5 [25]. Özorhan MO et al. have employed a hybrid approach based on SVM and GA (Genetic Algorithm) to predict the best currency pair for exchange. Authors have used primary technical indicators for their analysis and found that by mixing the raw data with a technical financial indicator, one is able to achieve more accurate results [26]. Khedr et al. have predicted the stock value using news sentiment analysis. Authors classified the results of Yahoo, Microsoft, and Facebook using three different approaches, namely, K-NN, SVM, and naïve Bayes [27]. Desai and Gandhi have designed a natural language processing (NLP) module for stock forecasting that uses the online news to determine the future stock value. The NLP was employed to find the polarity of sentences [28]. Zhao and Wang have used an outlier data mining technique for stock forecasting. Authors tried to remove the anomalies of the time series approach. Authors found that their method generated better long-term forecasting results for the Chinesemarket [29]. Bini and Mathew have used clustering and multiple regression techniques for stock forecasting. The objective of clustering is to find a set of companies where a customer has to invest money in better results. Different indexes like Jaccard, C, Rand, and Silhouette were used to validate the results. In general, the focus was on technical analysis, classification, and prediction only [30]. Huang and Gang have devised a kernel manifold learning approach for financial dataset analysis. They found their approach to be useful in improving accuracy. Moreover, the objective criterions provided by the kernel manifold learning approach also assist in depicting and predicting the precise volatility of the stock market [31]. Ye and Li have reviewed literature related to the role of big data in the capital market. Authors concluded that internet big data plays a significant role in stock analysis using sentiment analysis. Authors did not found any clear evidence that explicitly supports that the capital market can be predicted using internet big data [32]. M. Khashei, Z. Hajirahimi has examined the performance of series and parallel strategies in forecasting financial time series. Authors found that the hybridization of a multilayer perceptron model along with ARIMA produces better results when compared with those of the individual models [33]. Nayak and Misra have employed a GA-weighted condensed polynomial neural network (GACPNN). Authors applied GACPNN for five different stock indexes, namely, BSE, DJIA, NASDAQ, FTSE, and TAIEX. The model was validated using the Deibold–Mariano test and found to produce more accurate results [34].

3. Methodology

Statistics represent a multidisciplinary data exploration approach that has been effectively used in various fields such as engineering, physics, chemistry, economics, finance, commerce, computer science, and so on [35–39]. The effective use of different statistical techniques can help in examining the nature, distribution, and trends of data. Descriptive and inferential techniques are significant classes of statistical techniques. Descriptive techniques are aimed at providing aggregated information, that is,

they analyze the average and dispersion of data. However, they do not attempt to describe the nature of the population from which the sample has been taken. Rather, they examine the distribution of data. A measure of central tendency, dispersion, skewness, kurtosis, and correlation study are some of the conventional standards of descriptive statistics. Inferential statistics come into picture when one has to analyze the massive amount of data consisting of population size or an order of millions, billions, or even more. With this size of population, it is not feasible to acquire the data for each item of a population. Thus, inferential statistics are used to disclose the nature of the entire population using a sample from the population. Estimation statistics and hypothesis testing are common methods of inferential statistics [40].

Supervised learning (classification) is an important data mining technique that assists in categorizing data into important classes. Classification is a learning process in which a function F_n tries to map each instance of data set to tone specific classes. There are two types of classification techniques, namely descriptive and predictive. Descriptive modeling assists in finding a set of features that can effectively be used to recognize different classes, whereas predictive modeling is used to forecast the unknown category of data instances. These are more efficient for binary or nominal classes and are not fit for ordinal classes [41]. Preceding research observed that many supervised learning techniques have been used to solve different data mining problems [5,6,41]. Some of the dominant classification techniques are briefly introduced in the remaining part of this section.

Naïve Bayes is a conditional probability-based classifier that is highly scalable and gives equal importance to all attributes of the classification problem [42]. C4.5 is a decision tree-based technique that employs a top-down recursive divide and conquer approach for data classification [43,44]. Random forest is an ensemble-based classifier that can be used for enormous and multifaceted databases for exploration, classification, and prediction [45,46]. SVM is one of the discriminative classifiers in which classification is based on the decision planes (multidimensional or hyperplanes) and their boundaries, and is effectively used for both classification and regression. On the basis of the kernel function, SVM can be further categorized as linear, quadratic, cubic, fine Gaussian, and medium Gaussian SVM [47,48]. Logistic regression is a binary classification technique that cannot be applied to a problem where there are more than two classes to be classified. It can provide best fit for real-life issues like spam detection, banking, health, and marketing related applications. Unlike logistic regression, linear discriminant analysis (LDA) is a statistical classifier that can be used for data classification problems where data have to be categorized into two or more classes [49,50].

In spite of classification techniques, ICICI Bank's stock data were also examined using distinct statistical measures of central tendency and dispersion. The skewness and kurtosis were analyzed to investigate the trend of different attributes of stock data along with the distribution of data. Additionally, A-squared values and *p*-values of different attributes of ICICI Bank's stock were also investigated. Furthermore, the data were classified using different classifiers, such asnaïve Bayes; C4.5; random forest; logistic regression; linear discriminant; and linear, quadratic, cubic, fine, and medium Gaussian SVM.

Data Set

The last twelve years' (2007 to 2018) ICICI Bank's stock data extracted from Yahoo finance were extensively analyzed using several statistical and supervised learning techniques. There were 2714 distinct instances, along with seven different attributes. For a precise analysis, the cases comprising missing values were eliminated, and finally, 2706 instances were analyzed. Here, status represents the intraday investment analysis. In a day, if closing value is higher than the opening value of ICICI Bank's stock, then it will be a profitable day for the investor. Otherwise, it will represent a loss for the investor. The data were examined from different statistical perspectives. Moreover, different classifiers have been employed to classify 2714 distinct instances, and their performance was examined on the basis of different parameters.

4. Results

Different descriptive statistical measures were examined to explore the nature of ICICI Bank's stock data. Table 1 depicts four essential attributes of ICICI Bank's stock data, along with the values of several descriptive statistical measures.

Table 1. Descriptive statistical measures of ICICI Bank's Stock.

Attributes	Mean	Std. Dev	Variance	Skewness	Kurtosis	*p*-Value	A-Squared	95% Confidence Interval (Mean)
Open	201.82	61.35	3763.33	0.09	−0.32	0.005	15.29	199.50–204.13
Close	201.54	61.32	3760.47	0.09	−0.31	0.005	15.31	199.23–203.85
Low	198.43	61.09	3731.52	0.08	−0.31	0.005	15.06	196.12–200.73
High	204.79	61.57	3790.44	0.1	−0.32	0.005	15.43	202.47–207.11

It was found that the average value of open, close, low, and high attributes of ICICI Bank's stock lies between 198.43 to 204.79. As per statistical results, the 95% confidence interval range (mean) for open attribute reveals that the average value of open attribute lies between 199.50 to 204.13. From standard deviation, it was found that in the last twelve years, 68% of opening values lie between 140.47 and 263.17, 95% lie between 79.12 and 324.52, and 99.7% lie between 17.77 and 285.87. The negative kurtosis values of open, close, low, and high attributes represent that the distribution curve is platy curtic and more flat in nature.

Figure 2 depicts the brief and consolidated picture of the major attributes of ICICI Bank's stock data. It is observed that over the last 12 years, the minimum and maximum values of opening and closing balance lie between 47.95 to 360.80 and 47.81 to 362.30, respectively. A significant variation (652.45%) in minimum and maximum opening values of ICICI Bank's stock has been witnessed.

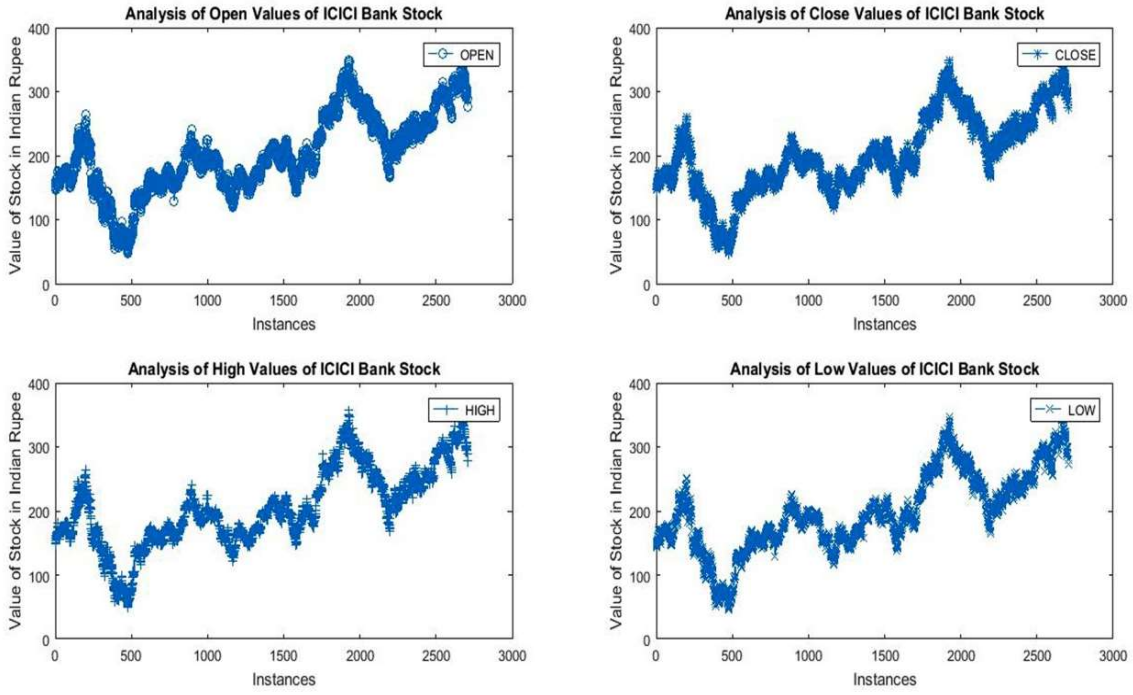

Figure 2. ICICI Bank's stock attributes.

From the data, it is also clear that maximum and minimum low–high variations are recorded in 2008 and 2013, respectively. However, the common range of variation lies between 90 and 120. Figure 3 depicts the daily percentage change recorded over the last 12 years. It is observed that in the years 2008, 2011, 2015, 2016, and 2017, large numbers of negative daily percentage changes were witnessed.

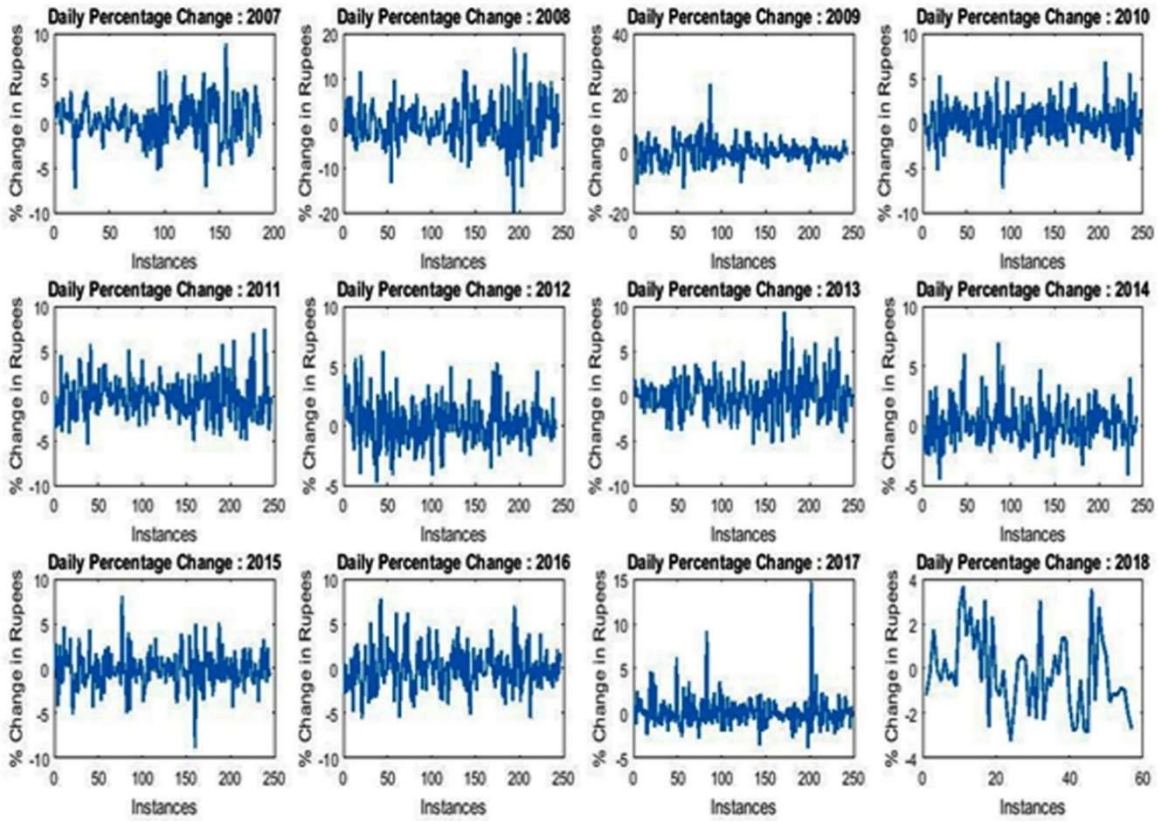

Figure 3. Daily percentage change.

Furthermore, in the last twelve years, the year 2009 seems to be very negative for the sellers as it had the lowest minimum and maximum opening values. However, the variation between minimum and maximum was found to be significant (264.2%). The year 2018 represents a sound state for investors who invested their money in previous years as the maximum opening value (Rs. 360.8) was achieved in this year. Additionally, the maximum and minimum variation in opening balance was recorded in the years 2008 and 2013, respectively. Figure 4 represents the summary of descriptive statistical measures for low–high variation of ICICI Bank's stock. The mean, minimum, and maximum values of high–low variation were found to be 6.36, 0, and 42.4, respectively. This means the maximum intraday variation in ICICI Bank's stock lies between 0 and 42.4 Indian rupees.

4.1. Evaluation Criterions and Analysis of Different Classifiers

The performance of predictive classification models is based upon the values of correctly and incorrectly classified instances. A confusion matrix represents the performance metrics of classifiers that highlight the number and types of errors made during data classification and are related to the following conditions:

- Positive instances classified as positive (TP)
- Positive instances classified as negative (FP)
- Negative instances classified as negative (TN)
- Negative instances classified as positive (FN)

Some metrics from the confusion matrix, such as accuracy, precision, recall, F1-score, specificity, and sensitivity, can be computed to determine the performance of classifiers from a different perspective.

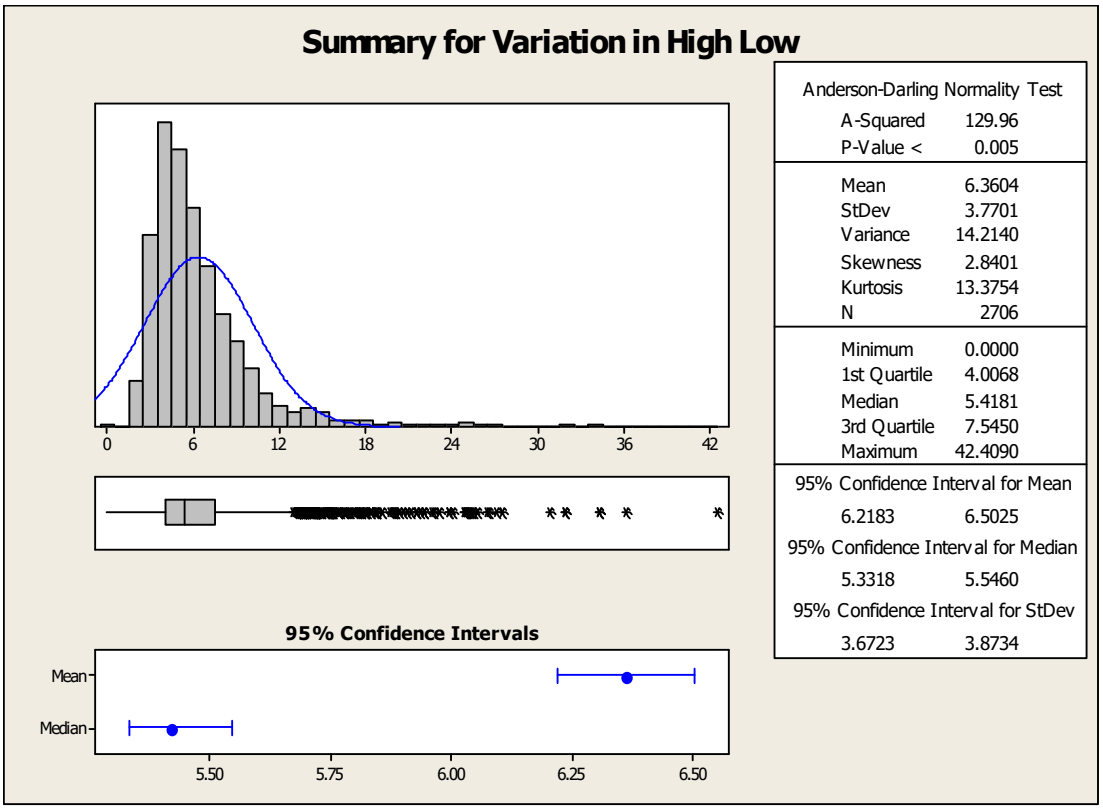

Figure 4. Summary of variation in high–low value of ICICI Bank's stock.

Accuracy is the most instinctive performance metric that represents the ratio of correctly foretold observation to the total observations, that is,

$$\text{Accuracy} = (TP + TN)/TP + TN + FP + FN \quad (1)$$

The rate of misclassification is an important measure of classification techniques. The rate of misclassification is based upon three major parameters of classification matrix, namely, true positive, true negative, and a total number of instances. A classifier that has zero rates of misclassification would be perfect and preferred. However, because of the presence of noise in data, it is difficult to find such a type of classifier. Mathematically, the rate of misclassification, which is denoted as err, is computed as

$$\text{Err} = (FP + FN)/N \quad (2)$$

Here, TP, TN, and N represent true positive, true negative, and total number of instances, respectively.

Sensitivity and specificity are computed to examine the rate of true positive and true negative instances. Mathematically,

$$\text{Sensitivity (TP Rate)} = TP/N \quad (3)$$

$$\text{Specificity (FP Rate)} = FP/N \quad (4)$$

Additionally, the precision and recall can be computed to determine the exactness and completeness property of the classifier.

$$\text{Precision} = TP/(TP + FP) \quad (5)$$

$$\text{Recall} = TP/(TP + FN) \quad (6)$$

Mean absolute error (MAE) represents the magnitude of the average absolute error. Mathematically,

$$MAE = \frac{1}{n\sum_{k=0}^{n}|e_k|} \quad (7)$$

F1 denotes the weighted average of recall and precision. It should be noted that a higher value of F1-score does not guarantee that the classifier is performing well. Rather, it depends upon the circumstances.

In this section, the performance of ten different classifiers has been examined in classifying the instance of ICICI Bank's stock data. Tables 2–11 representthe values of different metrics like FP, TP, TN, and FN, along with the number of correctly and incorrectly classified instances, accuracy, precision, recall, and an F1-score of different supervised classifiers in analyzing the data set of ICICI Bank's stock. From Tables 2–11, it is observed that rates of classification of naïve Bayes; C4.5; random forest; logistic regression; linear discriminant; and linear, quadratic, cubic, fine and medium Gaussian support vector machines lie between 47.6% and 48.9%, 53.6% and 53.6%, 53.0% and 53.6%, 99.7% and 99.8%, 53.6% and 53.6%, 98.7% and 99.8%, 91.0% and 93.9%, 75.0% and 91.7%, 78.2% and 79.9%, and 69.6% and 72.4%, respectively. To precisely examine the performance of different classifiers, the K-fold cross-validation mechanism was used. In K-fold validation, initially, the data have to be decomposed into K mutually exclusive equal sized folds or subsets. In 5-fold, the data are decomposed into giving subsets also known as folds (F1, F2, F3, F4, and F5). Testing and training are carried out five times. In the first iteration, the fold F1 acts as the test set and the remaining four subsets as training sets. Similarly, in the second iteration, F2 acts as testing and the remaining subgroups are used for drilling. The process is repeated five times. The data were mined by varying the folds from 5 to 10.

Table 2. Performance analysis of naïve Bayes.TP—true positive; TN—true negative; FP—false positive; FN—false negative.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	437	872	1018	387	1309	1405	48.23	30.03436
6	428	867	1027	392	1295	1419	47.71	29.41581
7	461	863	994	396	1324	1390	48.78	31.68385
8	455	865	1000	394	1320	1394	48.63	31.27148
9	445	849	1010	410	1294	1420	47.67	30.58419
10	457	872	998	387	1329	1385	48.96	31.40893

Table 3. Performance analysis of C4.5.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1455	0	0	1259	1455	1259	53.61091	100
6	1455	0	0	1259	1455	1259	53.61091	100
7	1455	0	0	1259	1455	1259	53.61091	100
8	1455	0	0	1259	1455	1259	53.61091	100
9	1455	0	0	1259	1455	1259	53.61091	100
10	1455	0	0	1259	1455	1259	53.61091	100

Table 4. Performance analysis of random forest.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1410	30	45	1229	1440	1274	53.05822	96.90722
6	1407	47	48	1212	1454	1260	53.57406	96.70103
7	1413	42	42	1217	1455	1259	53.61091	97.1134
8	1424	28	31	1231	1452	1259	53.50037	97.86942
9	1425	18	30	1241	1443	1259	53.16875	97.93814
10	1413	39	42	1220	1452	1259	53.50037	97.1134

Table 5. Performance analysis of logistic regression.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1444	1256	3	3	2700	6	99.77827	99.79267
6	1445	1258	2	1	2703	3	99.88914	99.86178
7	1442	1258	5	1	2700	6	99.77827	99.65446
8	1445	1257	2	2	2702	4	99.85218	99.86178
9	1445	1258	2	1	2703	3	99.88914	99.86178
10	1444	1256	3	3	2700	6	99.77827	99.79267

Table 6. Performanceanalysis of linear discriminant.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1074	381	366	893	1455	1274	53.61091	74.58333
6	1070	385	365	894	1455	1279	53.61091	74.56446
7	1089	366	381	878	1455	1244	53.61091	74.08163
8	1062	393	368	891	1455	1284	53.61091	74.26573
9	1080	375	369	890	1455	1265	53.61091	74.53416
10	1072	383	375	884	1455	1267	53.61091	74.08431

Table 7. Performance analysis of linear support vector machine (SVM).

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1437	1253	18	6	2690	24	99.1157	98.76289
6	1427	1252	28	7	2679	35	98.71039	98.0756
7	1436	1253	19	6	2689	25	99.07885	98.69416
8	1431	1252	24	7	2683	31	98.85777	98.35052
9	1437	1250	18	9	2687	27	99.00516	98.76289
10	1431	1256	24	3	2687	27	99.00516	98.35052

Table 8. Performance analysis of quadratic SVM.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1377	1172	78	87	2549	165	93.92041	94.63918
6	1346	1168	109	91	2514	200	92.6308	92.50859
7	1362	1184	93	75	2546	168	93.80987	93.60825
8	1403	1082	52	177	2485	229	91.56227	96.42612
9	1360	1133	95	126	2493	221	91.85704	93.47079
10	1349	1121	138	106	2470	244	91.00958	90.71957

Table 9. Performance analysis of cubic SVM.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	967	1070	488	189	2037	677	75.05527	66.46048
6	1247	1201	208	58	2448	266	90.19897	85.70447
7	1383	1106	72	153	2489	225	91.70965	95.05155
8	1182	1187	273	72	2369	345	87.28814	81.23711
9	1245	953	210	306	2198	516	80.98747	85.56701
10	1234	1127	221	132	2361	353	86.99337	84.811

Table 10. Performance analysis of fine Gaussian SVM.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1246	877	209	382	2123	591	78.22402	85.63574
6	1244	907	211	352	2151	563	79.25571	85.49828
7	1244	892	211	367	2136	578	78.70302	85.49828
8	1254	917	201	342	2171	543	79.99263	86.18557
9	1261	909	194	350	2170	544	79.95578	86.66667
10	1254	913	201	346	2167	547	79.84525	86.18557

Table 11. Performance analysis of medium Gaussian SVM.

Folds	TP	TN	FP	FN	Correctly Classified	Incorrectly Classified	Accuracy	Precision
5	1383	515	72	744	1898	816	69.93368	95.05155
6	1353	538	102	721	1891	823	69.67576	92.98969
7	1363	563	92	696	1926	788	70.96536	93.67698
8	1375	582	80	677	1957	757	72.10759	94.50172
9	1378	587	77	672	1965	749	72.40236	94.7079
10	1371	593	84	666	1964	750	72.36551	94.2268

The experimental analysis shows that logistic regression, followed by linear SVM, was found to be best suited as a classifier for ICICI Bank's stock analysis. NB, C4.5, RF, LD, and CSVM merely act as a random guessing machine. The rate of accuracy achieved using logistic regression lies in between 99.7% and 99.8%. Moreover, this classifier had a higher rate of precision, as well as recall. It was found that naïve Bayes seems to merely guessing machine, as it has the lowest rate of accuracy among all the classifiers. The rate of classification in classifying correct and incorrect instances using naïve Bayes was found to be 47.6% and 48.9%, respectively. In addition, when precision was considered, C4.5 seemed to be the best classifier.

The value of precision accomplished using six to ten cross-fold remained constant and, surprisingly, was 100%. However, this classifier utterly failed to classify true negative cases. Additionally, the rate of identifying false negative cases using naïve Bayes was extremely high. Like accuracy, logistic regression also showed outstanding performance as far as F1 values were concerned.

The rates of correctly and incorrectly classified instances achieved using different classifiers are depicted in Figure 5.

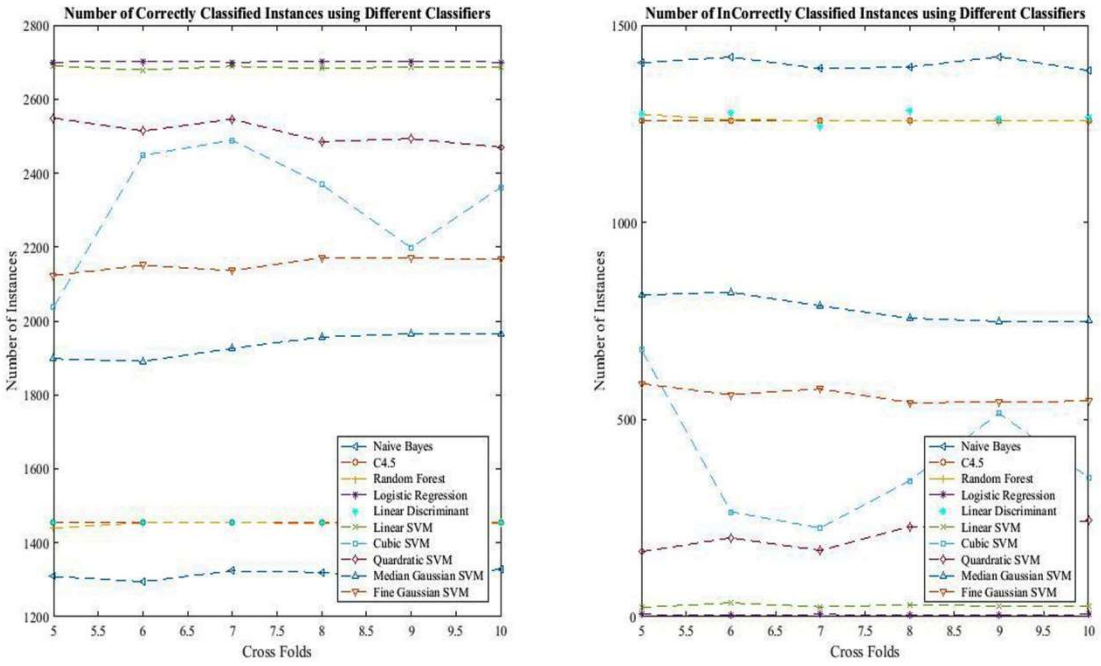

Figure 5. Number of correctly classified instances using different classifiers.

Figure 6 depicts the rate of misclassification of different classifiers. It is observed that naïve Bayes had the highest rate of misclassification, whereas logistic regression and linear SVM were found to have the lowest misclassification rate.

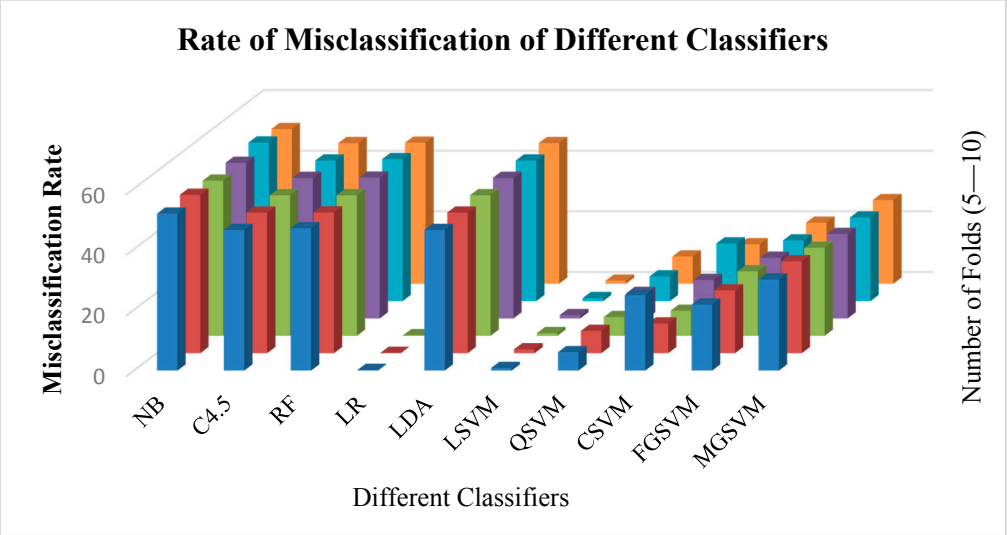

Figure 6. Number of correctly classified instances using different classifiers. NB—naïve bayes; RF—random forest; LR—logistic regression; LDA—linear discriminant analysis; LSVM—linear support vector machine; QSVM—quadratic SVM; CSVM—cubic SVM; FGSVM—fine Gaussian SVM; MGSVM—medium GSVM.

Additinally, the analysis of sensitivty, specificity, F1 score and recall has been presntedin Figure 7.

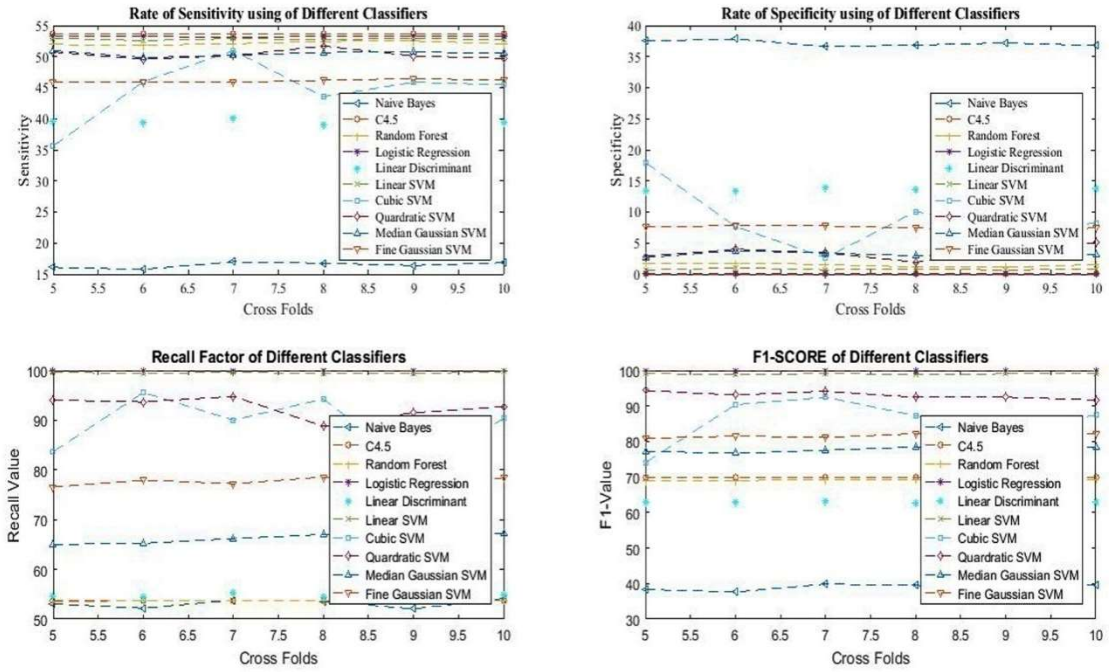

Figure 7. Analysis of sensitivity, specificity, F1, and recall of different classifiers.

4.2. Ranking Using MCD Techniques

Table 12 represents the summarized performance of different classifiers in classifying ICICI Bank's stock data. It represents a multi-criterion decision problem with ten different approaches having seven different performance criterions. This table has been created from Table 4 by taking the best possible value from different cross folds.

Table 12. Summarized results of Tables 2–11. LOG REG—logistic regression; LD—linear discriminant; LSVM—linear SVM; QSVM—quadratic SVM; CSVM—cubic SVM; FGSVM—fine Gaussian SVM; MGSVM—medium GSVM.

Mining Approach	Accuracy	Precision	Misclassification Rate	Sensitivity	Specificity	Recall	F1 Score
Naïve bayes	48.96831	31.68385	52.32	16.99	37.84	54.14692	39.87889
C4.5	53.61091	100	46.39	53.61	0	53.61091	69.80091
Random Forest	53.61091	97.93814	46.94	52.51	1.77	53.72624	69.2944
LOG REG	99.88914	99.86178	0.22	53.24	0.18	99.93084	99.8963
LD	53.61091	74.58333	46.39	40.13	14.04	55.3635	63.36922
LSVM	99.1157	98.76289	1.29	52.95	1.03	99.79079	99.17184
QSVM	93.92041	96.42612	8.99	51.69	5.08	94.78079	94.34738
CSVM	91.70965	95.05155	24.94	50.96	17.98	95.55556	92.47743
FGSVM	79.99263	86.66667	21.78	46.46	7.77	78.57143	82.25701
MGSVM	72.40236	95.05155	30.32	50.96	3.76	67.30486	78.63053
WEIGHT	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286	0.14286

This MCD problem has been solved using two different statistical techniques, called TOPSIS (technique for order preference by similarity to ideal solution) and WSA (weighted sum approach) [51,52]. The ranking of different approaches based upon seven different criterions is presented in Table 13.

Table 13. Ranking of different classification approaches using the weighted sum approach (WSA) and the technique for order preference by similarity to ideal solution (TOPSIS).

Approach	WSA	TOPSIS
Logistic Regression	1	1
Linear SVM	2	2
Quadratic SVM	3	3
Cubic SVM	4	5
Fine Gaussian SVM	5	4
Medium Gaussian SVM	6	6
C4.5	7	7
Random Forest	8	8
Linear Discriminant	9	9
Naïve Bayes	10	10

The working of the WSA method is based on the utility maximization principle. It helps in finding the ranks of the alternatives on the basis of their total utility by considering all the chosen criteria. In TOPSIS, d_i^+ and d_i^- represent the distance of ideal and basal variants. Here, *Hj* and *Dj* are the maximum or minimum values corresponding to the ideal or basal distances.

$$di^+ = \sqrt{\sum_{j=1}^{r} (Wij - Hj)^2} \tag{8}$$

$$di^- = \sqrt{\sum_{j=1}^{r} (Wij - Dj)^2} \tag{9}$$

Finally, the relative closeness to the ideal solution C_i is calculated as mentioned below:

$$C_i = d_i^- / (d_i^+) + (d_i^-) \tag{10}$$

In order to get the real picture of predicted rate of return, the ICICI stock data were also predicted using linear and multiple regression. Table 14 represents the difference between the rate of actual and predicted return value obtained using both linear and multiple regression. Here, the rate of return was computed for the month of February 2018. The buy-and-hold time was fixed at one month.

Table 14. Difference between actual and predicted rate of return.

Multiple Regression	Linear Regression
2.71	−0.01
0.09	−0.01
0.84	−0.01
7.14	0.00
0.09	−0.01
0.27	−0.01
0.84	0.00
−0.03	−0.01
0.24	−0.01
0.02	0.00
−2.03	0.00
−3.27	0.00
−1.79	0.00
−0.34	0.01
−1.31	0.01
1.33	0.01
−2.85	0.01
−0.35	0.01
−2.55	0.01

From Table 14, it was found that the results obtained using linear regression were more precise when compared with results obtained using multiple regression, as the difference between actual and predicted rate of return was very small for linear regression.

5. Conclusions

ICICI Bank's stock was substantially examined using different statistical and supervised learning techniques. The large negative variation observed in five years (2008, 2011, 2015, 2016, and 2017) indicates that in these years, a momentous intraday loss was recorded. A negatively skewed representation indicates that the distribution curve is platy curtic and more flat in nature. The lowest minimum and maximum opening values were marked in 2009. Therefore, it seems that the long-term investors who invested their money in this year must have achieved a good rate of return. The year 2018 represented a sound state of ICICI Bank's stock as the maximum opening value of Rs. 360.8 was achieved in this year. Therefore, this year should not be seen as a year of investment. This study can be extended to predict the daily, weekly, and monthly future values of ICICI Bank's stock. Furthermore, it was observed that rates of classification of naïve Bayes; C4.5; random forest; logistic regression; linear discriminant; and linear, quadratic, cubic, fine, and medium Gaussian support vector machines lie between 47.6% and 48.9%, 53.6% and 53.6%, 53.0% and 53.6%, 99.7% and 99.8%, 53.6% and 53.6%, 98.7% and 99.8%, 91.0% and 93.9%, 75.0% and 91.7%, 78.2% and 79.9%, and 69.6% and 72.4%, respectively. The performance of logistic regression was outstanding when compared with other classifiers and this was validated using two different multi-criterion decision problem techniques, namely TOPSIS and WSA The rank generated using TOPSIS and WSA verified the outstanding performance of logistic regression. In addition to this, the average values of major attribute (open, close, low, and high) lie between 198.43 to 204.79. Moreover, based upon the performance of difference classifiers, an innovative and novel ensemble-based classifier can be designed. In linear and multiple regression, as far as the rate of return is concerned, the results produced using linear regression are better than the results obtained using multiple regression.

Author Contributions: Conceptualization, M.S. and S.S.; Methodology, M.S., S.S.; Software, M.S.; Validation, M.S., G.S.; Formal Analysis, M.S.; Investigation, M.S., S.S. and G.S.; Resources, M.S.; Data Curation, M.S.; Writing-Original Draft Preparation, M.S.; Writing-Review & Editing, M.S., S.S. and G.S; Visualization, M.S.; Supervision, G.S.; Project Administration, M.S., G.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sharma, R. ICICI Bank Equity Research. 22 January 2016. Available online: https://www.sanasecurities.com/icici-bank-equity-research (accessed on 20 September 2018).
2. IANS. SBI India's Most Trusted Bank, ICICI Top in Private Sector: Report. 19 April 2018. Available online: https://economictimes.indiatimes.com/industry/banking/finance/banking/sbi-indias-most-trusted-bank-icici-tops-in-private-sector-report/articleshow/63818576.cms (accessed on 20 September 2018).
3. Tsai, C.; Lai, C.; Chao, H.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**, *2*, 1–32. [CrossRef]
4. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.* **2007**, *32*, 995–1003. [CrossRef]
5. Han, J.; Kamber, M.; Pei, J. *Data Mining Concepts and Techniques*; Morgan Kauffmann Publishers: Burlington, MA, USA, 2015.
6. Kaur, P.; Sharma, M. Analysis of Data Mining and Soft Computing Techniques in Prospecting Diabetes Disorder in Human Beings: A Review. *Int. J. Pharm. Sci. Res.* **2018**, *9*, 2700–2719.
7. Rajesh, D. Application of spatial data mining for agriculture. *Int. J. Comput. Appl.* **2011**, *15*, 7–9. [CrossRef]
8. Bhargavi, P.; Jyothi, S. Applying naive Bayes data mining technique for classification of agricultural land soils. *Int. J. Comput. Sci. Netw. Secur.* **2009**, *9*, 117–122.
9. Liao, S.-H.; Chu, P.-H.; Hsiao, P.-Y. Data mining techniques and applications—A decade review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]
10. Kadam, S.; Raval, M. Data Mining in Finance. *Int. J. Eng. Trends Technol.* **2014**, *16*, 377–381. [CrossRef]
11. Enke, D.; Thawornwong, S. The use of data mining and neural networks for forecasting stock market returns. *Expert Syst. Appl.* **2005**, *29*, 927–940. [CrossRef]
12. Fu, T.-C. A review on time series data mining. *Eng. Appl. Artif. Intell.* **2011**, *24*, 164–181. [CrossRef]
13. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.F.; Hua, L. Data mining in healthcare and biomedicine: A survey of the literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef] [PubMed]
14. Huang, M.-J.; Chen, M.-Y.; Lee, S.-C. Integrating data mining with the case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Syst. Appl.* **2007**, *32*, 856–867. [CrossRef]
15. Aljumah, A.A.; Ahamad, M.G.; Siddiqui, M.K. Application of data mining: Diabetes health care in young and old patients. *J. King Saud Univ.-Comput. Inf. Sci.* **2013**, *25*, 127–136. [CrossRef]
16. Sharma, M.; Singh, G.; Singh, R. Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. *IRBM* **2017**, *38*, 305–324. [CrossRef]
17. Sharma, M.; Singh, G.; Singh, R. An Advanced Conceptual Diagnostic Healthcare Framework for Diabetes and Cardiovascular Disorders. *EAI Endorsed Trans. Scalable Inf. Syst.* **2018**, *5*, 1–11. [CrossRef]
18. Yang, Y.; Adelstein, S.J.; Kassis, A.I. Target discovery from data mining approaches. *Drug Discov. Today* **2012**, *14*, 147–154. [CrossRef] [PubMed]
19. Ananiadou, S.; Kell, D.B.; Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol.* **2006**, *24*, 571–579. [CrossRef] [PubMed]
20. Chandralekha, M.; Shenbagavadivu, N. Performance Analysis of Various Machine Learning Techniques to Predict Cardiovascular Disease: An Emprical Study. *Appl. Math. Inf. Sci.* **2018**, *12*, 217–226. [CrossRef]
21. Manjula, C.; Muniyal, B.B. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. *Procedia Comput. Sci.* **2016**, *89*, 117–123.
22. Sangeeta, M. ICICI Bank: A Multivariate Analysis of Customer's Acceptability. *Glob. J. Manag. Bus. Res.* **2011**, *11*, 1–9.
23. Pooja, R. A Study of Financial Performance: A Comparative Analysis of AXIS and ICICI Bank. *Int. J. Multidiscipl. Res. Dev.* **2017**, *4*, 12–20.
24. Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Syst. Appl.* **2015**, *42*, 259–268. [CrossRef]

25. Al-Radaideh, Q.I.; Assaf, A.A.; Alnagi, E. Predicting Stock Price Using Data Mining Technique. In Proceedings of the International Arab Conference on Information Technology (ACIT'2013), Katumu, Sudan, 17–19 December 2013; pp. 1–8.
26. Özorhan, M.O.; Toroslu, I.H.; Tolga, O.; Glu, S. A strength-biased prediction model for forecasting exchange rates using support vector machines and genetic algorithms. *Soft Comput.* **2017**, *21*, 6653–6671. [CrossRef]
27. Khedr, A.E.; Salama, S.E.; Yaseen, N. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *Int. J. Intell. Syst. Appl.* **2017**, *7*, 22–30.
28. Desai, R.; Gandhi, S. Stock Market Prediction Using Data Mining. *Int. J. Eng. Dev. Res.* **2014**, *2*, 2780–2784.
29. Zhao, L.; Wang, L. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm. In Proceedings of the IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 26–28 August 2015.
30. Bini, B.S.; Mathew, T. Clustering and Regression Techniques for Stock Prediction. *Procedia Technol.* **2016**, *24*, 1248–1255. [CrossRef]
31. Huang, Y.; Gang, K. A kernel entropy manifold learning approach for financial data analysis. *Decis. Support Syst.* **2014**, *64*, 31–42. [CrossRef]
32. Ye, M.; Li, G. Internet big data and capital markets: A literature review. *Financ. Innov.* **2017**, *3*, 1–18. [CrossRef]
33. Khashei, M.; Hajirahimi, Z. Performance evaluation of series and parallel strategies for financial time series forecasting. *Financ. Innov.* **2017**, *3*, 1–24. [CrossRef]
34. Nayak, S.C.; Misra, B.B. Estimating stock closing indices using a GA-weighted condensed polynomial neural network. *Financ. Innov.* **2018**, *4*, 1–21. [CrossRef]
35. Yan, F.; Robert, M.; Li, Y. Statistical methods and common problems in medical or biomedical science research. *Int. J. Physiol. Pathophysiol. Pharmacol.* **2017**, *9*, 157–163. [PubMed]
36. Du Prel, J.-B.; Röhrig, B.; Blettner, M. *Statistical Methods in Medical Research*; Deutsches Ärzteblatt International: Berlin, Germany, 2009; Volume 106, p. 99.
37. Zhan, W.; Fink, R.; Fang, A. Application of Statistics in Engineering Technology Programs. *Am. J. Eng. Educ.* **2010**, *1*, 65–78. [CrossRef]
38. Hamada, R.; Patell, J.M.; Staelin, R.; Wecker, W.E. The Role of Statistics in Accounting, Marketing, Finance, and Production. *J. Bus. Econ. Stat.* **1988**, *6*, 261–272.
39. Buenestado, P.; Acho, L. Image Segmentation Based on statistical confidence Intervals. *Entropy* **2018**, *20*, 46. [CrossRef]
40. Gillian, B. A Statistical Primer: Understanding Descriptive and Inferential Statistics. *Evid. Based Lib. Inf. Pract.* **2007**, *2*, 32–47.
41. Du, H. *Data Mining Techniques and Applications—An Introduction*, 1st ed.; Cengage Learning: Delhi, India, 2013.
42. Angelo, G.D.; Rampone, S.; Palmieri, F. Developing a trust model for pervasive computing based on Apriori association rules learning and Bayesian classification. *Soft Comput.* **2017**, *21*, 6297–6315.
43. Lin, S.W.; Chen, S.C. Parameter determination and feature selection for the C4.5 algorithm using scatter search approach. *Soft Comput.* **2011**, *16*, 63–75. [CrossRef]
44. Sharma, S.; Aggarwal, J.; Sharma, S. Classification through Machine Learning Technique: C4.5 Algorithm based on Various Entropies. *Int. J. Comput. Appl.* **2013**, *82*, 20–27.
45. Maragoudakis, M.; Serpanos, D. Towards Stock Market Data Mining Using Enriched Random Forests from Textual Resources and Technical Indicators. *IFIP Adv. Inf. Commun. Technol.* **2010**, *339*, 278–286.
46. Chen, F.H.; Howard, H. An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree. *Soft Comput.* **2015**, *20*, 1945–1960. [CrossRef]
47. Huang, W.; Nakamori, Y.; Wang, S.Y. Forecasting stock market movement direction with support vector machine. *Comput. Oper. Res.* **2005**, *32*, 2513–2522. [CrossRef]
48. Larose, D.T.; Larose, C.D. *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed.; Wiley Publishers: Delhi, India, 2016.
49. Banu, G.R. Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique. *Commun. Appl. Electron. (CAE)* **2016**, *4*, 4–6.

50. Maroco, J.; Silva, D.; Rodrigues, A.; Guerreiro, M.; Santana, I.; de Mendonça, A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of LDA, logistic regression, neural networks, SVM, classification trees and random forests. *BMC Res. Notes* **2011**, *4*, 1–14. [CrossRef] [PubMed]
51. Krohling, R.A.; Pacheco, A.G.C. A-TOPSIS—An Approach Based on TOPSIS for Ranking Evolutionary Algorithms. *Procedia Comput. Sci.* **2015**, *55*, 308–317. [CrossRef]
52. Kolios, A.; Mytilinou, V.; Lozano-Minguez, E.; Salonitis, K.A. Comparative Study of Multiple-Criteria Decision-Making Methods under Stochastic Inputs. *Energies* **2016**, *9*, 566. [CrossRef]