World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Stock feature dimensionality reduction for closing price prediction using unsupervised machine learning technique (case study of Nigeria Stock Exchange)

Mbeledogu Njideka Nkemdilim [1,*], Paul Roseline Uzoamaka [1], Ugoh Daniel [1] and Mbeledogu Kaodilichukwu Chidi [2]

[1] *Department of Computer Science, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria.*
[2] *Department of Mechanical Engineering, Faculty of Engineering, University of Ottawa, Canada.*

World Journal of Advanced Research and Reviews, 2024, 21(03), 1930–1936

Publication history: Received on 29 January 2024; revised on 21 March 2024; accepted on 23 March 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.3.0912

Abstract

Stock data offers invaluable insights into the world of finance. It encourages investment and savings for both individuals and the nation as a whole. To monitor and predict stock, many stock variables are collected which in turn leads to curse of dimensionality because they occupy much storage space and take more computational time. In order to avoid this, there is a need to reduce the dimensionality of the stock features. Since stock is an unlabeled data with a Gaussian distribution (the features are normally jointly distributed), an unsupervised machine learning technique was applied to discover, establish an association and extract the most important features that have the entire generality of the original dataset for predicting the next day's closing price. The dataset (daily price list) of Dangote Sugar Refinery Plc was randomly selected from the 27 blue chip companies in Nigeria Stock Exchange. 4 stock features were discovered and extracted from the 9 features in the original dataset.

Keywords: Curse of dimensionality; Unsupervised machine learning technique; Principal Component Analysis; Gaussian distribution

1. Introduction

Closing Price is the final price at which a security is traded on a given trading day. It gives the most up-to-date valuation of a security until trading commences again on the next trading day (Investopedia, 2013) as well as determines the performance of a particular stock. It is important to investors, financial institutions and other stakeholders because it provides a useful marker to assess changes in stock prices over time and also measure market sentiment for a given security over a trading day. In order to predict the closing price of any stock, stock features are required. These stock features can be dependent or independent variables that have complexities and are collectively referred to as dimensions.

Dimensions are the dependent and independent features or attributes that are present in a dataset. In the light of classification or regression tasks, a large volume of data is required to work with. If the large data consists large number of variables (features), it creates a very large window of data because big data yields sparsity which in turn results to Curse of Dimensionality. Sriram (2023) analyzed curse of dimensionality as the number of dimensions (features) increases, the amount of data needed to generalize the machine learning model accurately increases exponentially. A large data with sparse features causes increase in space and computational complexity, poor model performance and risk of overfitting. In order to avoid this, there is a need to reduce these features to a smaller number that best gives the generality of the entire features by refining and adjusting the dataset for the actual iterative algorithm.

* Corresponding author: Mbeledogu Njideka Nkemdilim

For a system to be termed intelligent, the system must be capable of reasoning and learning. Learning being an attribute of a human being is the process that leads to change which occurs as a result of experience and increases the potential for improved performance and future learning (Ambrose *et al*., 2010). The ability of machines to exhibit this human learning attribute is called Machine learning (ML). When machines learn under supervision, it is termed supervised learning. It is easier for a system to learn under supervision but when the decision boundary is overstrained and the training set does not have labels of the required predictions then it is needful to adopt the learning capabilities of machines when not under supervision.

In contrast to supervised, unsupervised learning does not require correct answers associated with each input pattern in the training dataset, that is, the labels of the data are not known. This learning paradigm discovers hidden patterns in the data by identifying groups of samples with specific characteristics that are similar. The measure of its accuracy is the likelihood of the selected model to discover certain data set (cluster or establish an association) based on the specified distribution (Hadžiabdić and Peters, 2021). It explores the underlying structure in the data and organizes patterns into groups from their correlations (Figure 1).It is also termed competitive learning. The major tasks of unsupervised learning are Clustering and Dimensionality Reduction (Mikalsen, 2018).

Clustering is the process of identifying natural groupings within multidimensional data according to some similarities based on the Euclidean distance measure (Omran *et al*., 2007) while dimensionality reduction on the other hand is the process of decreasing the dimensionality of the features (Zebari *et al*., 2020).

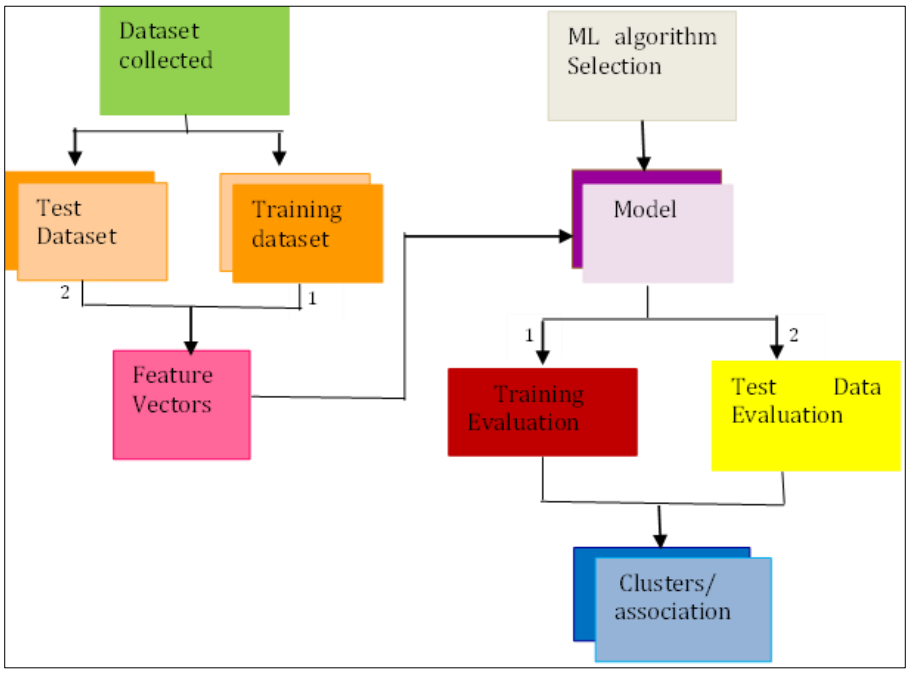

Figure 1 DFD of Unsupervised Machine Learning Paradigm

2. Literature review

Dimensionality reduction is the process of mapping a data sample from a high-dimensional space to a relatively low-dimensional space (Jia *et al*., 2022). It takes into account the correlation between several independent variables in a model (multi-collinearity) to remove noise, redundant features and dependent variables, reduce computational space and time, understand the visualization of data and interpret it (Barla, 2023). In this research, four techniques for dimensionality reduction will be discussed. They are Linear Discriminant Analysis (LDA), Independent Component Analysis, Non-negative Matrix Factorization and Principal Component Analysis (PCA).

2.1. Linear Discriminant Analysis (LDA)

This is a supervised machine learning technique for reducing dimension in a dataset that is linear and has normal distribution in nature. It aims at finding the linear discriminants to represent the axes that maximize separation between different classes of labeled data.

2.2. Independent Component Analysis (IDA)

This is a non-parametric unsupervised machine learning technique that assumes that data is determined by independent factors, uses linear method to reduce dimension and transforms the dataset into columns of independent non-Gaussian components. A set of vectors are transformed into maximally independent sources (Talebi, 2021) which are extracted in the second-order and higher-order correlations (Yinglin, 2020). It does not lay must emphasis on the change of the components reciprocal without observing its effect.

2.3. Non-Negative Matrix Factorization (NNMF)

NNMF is a linear unsupervised machine learning technique that determines only non-negative features by reducing a matrix with only non-negative coefficient into products of two other non-negative matrices with reduced ranks (Lopes and Ribeiro, 2015). It has the capacity to extract sparse and easily interpretable factors and thus, used in image processing, text mining and hyperspectral imaging (Coyler, 2019). Though it handles dimensionality reduction as it concerns unlabeled data, the initialization of the factorization is very important because poor initialization produces poor results. NNMF does not guarantee best factorization and can result to local minima.

2.4. Principal Component Analysis (PCA)

PCA is an unsupervised machine learning technique that reduces high dimensioned multivariate data features into essential features that are uncorrelated in nature. These essential features are linear combination of the initial variables and are called principal components (PCs). In order to achieve this, either correlation or covariance matrix is used to reduce the dimensionality of the data. When variables have non-identical scales, correlation matrix is used while covariance matrix is used for same scales.

PCA shows the relationship between observation and variables, and finds a subspace that conserves the variance of the data. Variance aids prediction or decision making as it depicts the distribution and variability of a set of data. Also, it resolves the problem of multicollinearity which causes skewedness or misleading of results, reduction of the precision of the estimated coefficients and increases the variance of the coefficient estimates.

3. Materials and method

Stock data has a Gaussian distribution (Owais, 2022; Amaral *et al*., 2000) and the effectiveness of the market should not be absolute and static but relative to show dynamic changes over time (Yang and Hou, 2022). Unsupervised principal component analysis machine learning paradigm will be employed for the reduction of the dimensionality of the stock price features for closing price prediction. This technique will capture the underlying structure in the data and organize patterns into groups from their correlations.

3.1. Data Collection

From the listed 27 blue chip companies in Nigeria Stock Exchange (companies with long record of stable and reliable stock growth), the historical stock data of three companies that are already household names in Nigeria were randomly selected as the research experimental data. Their daily price lists were captured from www.cashcraft.com for 2008-2011. These companies are: Dangote Sugar Refinery Plc (Food and Beverage), GlaxoSmith Kline Plc (Health) and Julius Berger Nig. Plc (Construction).

3.2. Stock Preprocessing

Minitab application package was used to implement the PCA and the Correlation matrix was adopted due to the nature of stock values (non-identical scales). The under listed categorized variables which are also known as the technical indicators on Table 1 are the quoted daily trading pricelist variables from the Nigeria Stock Exchange (NSE).

3.3. PCA's Operational Structure

A d-dimensional dataset comprising of n daily observations on a vector of n stock features/variables arranged in a matrix X (n x p) was collected:

$$\{X_1, X_2, \cdots, X_n\} \in R^p \qquad (1.0)$$

Where p are the stock variables/features as presented on Table 1.

Table 1 Categorized NSE quoted daily trading pricelist

Stock Variables (Features)	Component Number
Open Price	1
High Price	2
Low Price	3
Close Price	4
Change (Difference between open price and close price)	5
Chg (Difference between high price and low price)	6
Deals	7
Volume	8
Values	9

PCA steps for Dimensionality Reduction:

- Standardize the d-dimensional dataset.
- Construct the covariance matrix.
- Decompose the covariance matrix into its eigenvectors and eigenvalues.
- Sort the eigenvalues by decreasing order to rank the corresponding eigenvectors.
- Select k eigenvectors which correspond to the k largest eigenvalues, where k is the dimensionality of the of the new feature subspace($k \leq d$).
- Construct a projection matrix W from the "top" k eigenvectors.
- Transform the d-dimensional input dataset X using the projection matrix W to obtain the new k dimensional feature space.

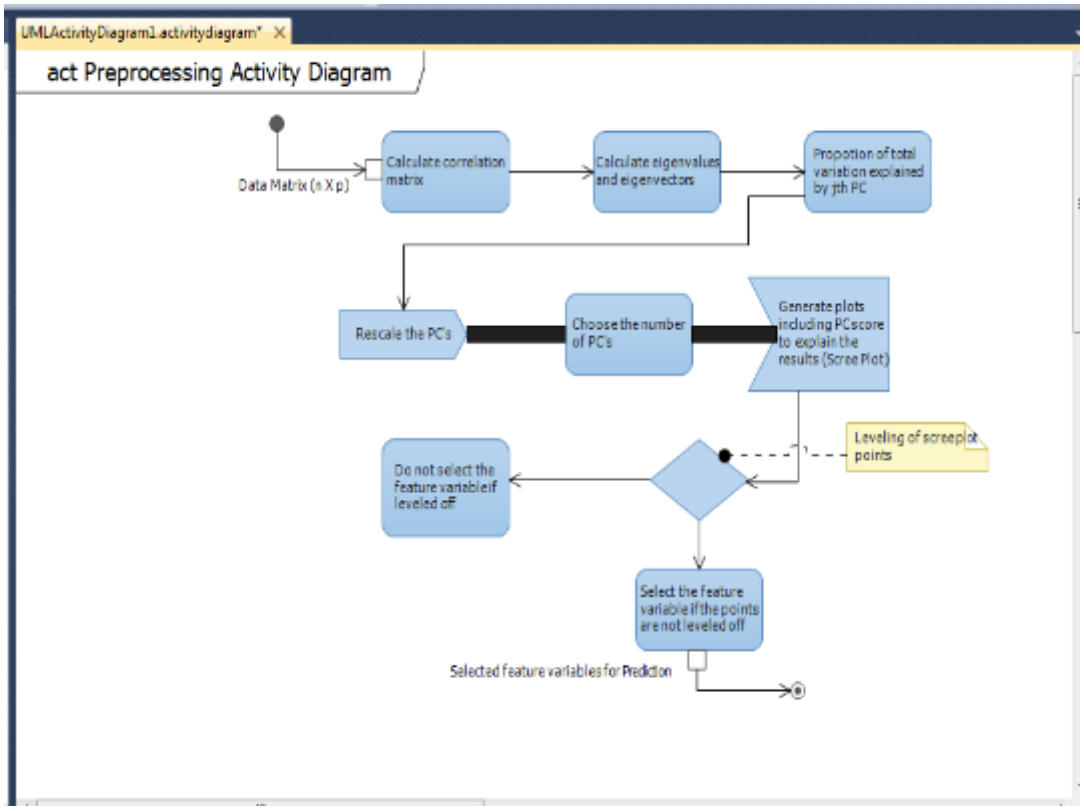

Figure 2 Activity diagram of the PCA

Figure 2 depicts the activity diagram of the PCA stock preprocessing. This shows the flow of control with emphasis on the sequence and conditions of the flow.

3.4. Extracted Stock Features

Table 2 and Figure 3 show the results of the PCA of the dataset and the scree plot of the analysis respectively. Scree plot is a simple line segment that shows a fraction of the total variance as represented by the PC. It graphically shows the optimal number of components to retain for further analysis.

Table 2 Principal Component Analysis (Dangote Sugar)

Eigen analysis of the Correlation Matrix									
Eigen value	5.1538	1.8692	1. 5743	0. 3578	0.0232	0.0085	0.0064	0.0050	0.0018
Proportion	0.573	0.208	0.175	0.040	0.003	0.001	0.001	0.001	0.000
Cumulative	0.573	0.780	0.955	0.955	0.998	0.999	0.999	1.000	1.000
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Open	0.410	0.207	-0.148	-0.125	0.706	-0.028	0.136	-0.173	-0.452
High	0.410	0.190	-0.184	-0.140	-0.118	-0.645	-0.153	-0.263	0.468
Low	0.412	0.216	0.134	-0.097	0.106	0.569	-0.096	0.421	0.485
Close	0.408	0.196	-0.190	-0.095	-0.689	0.160	0.185	-0.053	-0.461
Change	0.197	-0.579	-0.329	0.025	0.026	0.330	0.081	0.619	-0.135
Chg	0.141	-0.617	-0.341	0.027	0.009	0.352	-0.098	-0.575	0.133
Deals	-0.346	-0.001	-0.239	-0.906	-0.006	0.026	-0.019	0.033	-0.014
-0.014	-0.292	0.216	-0.534	0.232	0.038	0.008	0.688	-0.044	0.227
Value	-0.252	0.271	-0.568	0.265	0.005	0.005	-0.652	0.064	-0.203

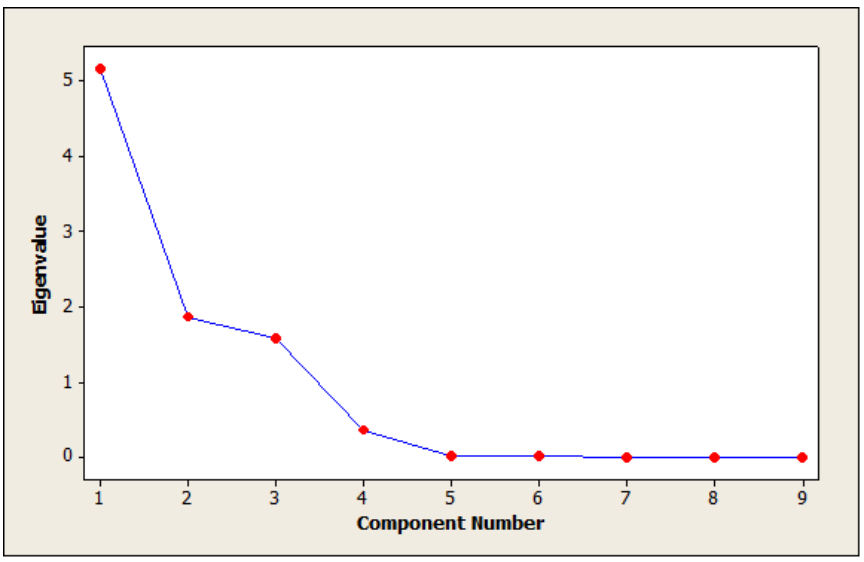

Figure 3 Scree Plot of the PCA Results and Discussion

From Table 2, each principal component was plotted against its eigenvalue. The analysis of the eigenvalue shows that variables 1, 2, 3, and 4 returned significant proportion greater than 0.10 while other variables had proportion less than 0.10. From the scree plot in Figure 3, Open price, High price, Low price and Close price had high eigenvalues and were

selected while the variables with principal components 5–9 that leveled off (the elbow) were not selected. These four extracted technical indicator are the stock inputs for predicting closing price. Table 3 with its stock dataset.

Table 3 Sample of Dangote Sugar Refinery Plc Dataset

Date	Open price	High Price	Low Price	Close Price
2/1/2008	40.001	40.101	38.501	40
3/1/2008	40.001	40.201	39.951	40.99
4/1/2008	40.991	41.001	39.501	41.89
7/1/2008	40.881	41.001	40.111	41.9
8/1/2008	41.891	41.911	40.701	42.99
9/1/2008	41.901	41.991	40.001	43.8
10/1/2008	42.991	43.001	40.011	43.9
11/1/2008	43.801	44.801	43.001	44.2
14/1/08	43.901	44.591	43.021	44.2
15/1/08	44.201	44.201	43.001	44
16/1/08	44.201	44.201	42.011	42.99
17/1/08	44.001	44.191	43.501	41.4
18/1/08	42.991	43.501	42.151	39.91
21/1/08	41.401	43.971	40.871	40
22/1/08	39.911	41.801	39.901	39.94
23/1/08	40.001	40.151	39.501	39
24/1/08	39.941	40.401	38.011	39.75

4. Conclusion

PCA serves as a good machine learning technique for data analysis exploration (dimensionality reduction). It transforms the dataset into uncorrelated (independent) PCs as it does not take into account the class labels in the data but aims to find the directions of maximum variance in the stock dataset.

Compliance with ethical standards

Disclosure of conflict of interest

There are no conflicts of interest.

References

[1] Investopedia (2013). Stock Market. Retrieved from www.investopedia.com/terms/s/stock market.asp#axzz2IjJu5E13

[2] Owais, .S. (2022). What is Normal Distribution in Financial Market. Retrieved from https://www.learnsignal.com/blog/what-is-normal-distribution/

[3] Amaral, N.F., Plerou, .V., Gopikrishnan, .P., Meyer, .M. and Stanley, .H.E. (2000). The distribution of returns of stock prices. International Journal of Theoretical and Applied Finance, Vol.3, No. 3, pgs.365-369

[4] Ambrose, S.A., Brides, .M.N., Dipietro, .M., Lovett, .M.C. and Norman, .M.K. (2010). How learning works: Seven Research-Based Principles for Smart Teaching, Published by Jossey-Bass A Wiley Imprint, San Francisco, Pgs. 1-301

[5] Sriram (2023). Curse of Dimensionality in Machine Learning: How to solve the Curse? Retrieved from https://www.upgrad.com/blog/curse-of-dimensionality-in-macine-learning-how-to-solve0-the-curse/

[6] Omran, .M., Engelbrecht, .A. and Salman, .A. (2017). An Overview of Clustering Methods. Intelligent Data Analysis 11(6): pgs. 583-605

[7] Zebari, .R., Abdulazeez, .A.M., Zeebaree, .D.Q., Zebari, .D.A. and Saeed, J.N. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction, Journal of Applied Science and Technology Trends, Vol.01, No. 02, pgs.56-70

[8] Jia, .W., Sun, .M., Lian, .J. and Hou, .S. (2022). Feature Dimensionaltiosmtionality Reduction: A review, Com,plex Intell. System, Vol. 8, pgs. 2663-2693

[9] Barla, .N. (2023). Dimensionality Reduction for Machine learni9ng. Retrieved from https://neptuine.ai/blog/dimensionalty-reduction

[10] Had*ž*iabdi*ć* . *K*. *K*. and Peters, .A. (2021). Chapter 15- Aritificial Integillience in Clinical decision-making for diagnosis of Cardiovascular disease using Epigenetics Mechanism. Epigenetics in Cardiovascular Disease, Vol. 24 in Translational Epigenetics. Pgs. 327-345

[11] Talebi. S. (2021). Independent Component Analysis (ICA) Finding hidden factors in data, Towards Data Science. Retrieved from https://towrdsdatascience.com/independent-component-analysis-ica-a3eba0ccec35

[12] Yinglin, .X. (2020). Correlation and Association analyses in microbiome study integrating multiomics in health and Disease, Progress in Molecular Biology and Translational Science, Science Direct, Vol. 171, Pgs. 309-491

[13] Lopes, .N. and Riberio, .B. (2015). Non-Negative Matrix Factorization (NMF). Machine learning for Adaptive Many-Core Machines- A Practical Approach Studies in Big Data, vol.7, Springer, Cham. https://doi.org/10.10071978-3-319-06938-8_7

[14] Mikalsen, .K.Ø. (2018). Advancing Unsupervised and Weakly Supervised Learning with Emphasis on Data-Driven Healthcare, A dissertation for the degree of Philosophiae Doctor, Faculty of Science and Technology, Department of Mathematics and Statistics. Pgs. 1-201

[15] Coyler, .A. (2019). The Why and How of non-negative matrix factorization. Retrieved from https://blog.acoyler.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/

[16] Yang, .P. and Hou, .X. (2022). Research on Dynamic Characteristics of Stock Market Based on Big Data Analysis, Discrete Dynamics in Nature and Society, Vol. 2022, Article ID 8758976, Pgs. 1-8