A Deep Learning Framework: Dimensionally Data Reduction and Predicting Intellegence Business Decision

Imtiage Ahmed, Probal Kumar Halder, K.M. Samiun Alim, Md. Ashiqur Rahman

Abstract— Business industries are seeking and taking a lot of decisions at a time. These decisions are taken by the management mostly depending on their experiences and previous business data. Business intelligence (BI) is a technology-driven process for analyzing data and presenting meaningful information to help the management to make better business decisions. The aim of this research is the development of a BI strategy in accordance with strategic business goals. We built a deep learning framework which reduces dimensions of the data as well as predicts intelligent business decisions based on historical data. We worked on over six thousand cases of a sales record. Two deep learning methods, principal component analysis and liner regression has been applied to reduce data dimensions reduction for visualization and predict intelligent decision.

Index Terms— Data reduction, Data visualization, Decision Prediction, Deep learning, Dimension reduction, Intelligent decision, Linear Regression , Principal Component Analysis.

1 INTRODUCTION

IN this twenty-first century, industries are focusing on their business data for better business decisions. From the customer's behavior to transactions, business industries are relying on each and every commercial data that is being produced. Many companies are still relying on the manual managing systems for which getting access to clean, high-quality data remains difficult for those companies. It is also making the managerial figures to face the challenges to take better commercial decisions under various circumstances.

Business Intelligence can provide accurate decisions from historical (all parts of the systematic money generated information and client information) data and can give real-time updates that are settled on reality-based choices but not just guesswork. Using business intelligence, information can be brought together and can be seen in a dashboard or report sparing tremendous measures of time and taking out wasteful aspects. That enables organizations to accurately identify current performance levels through data analysis, allows setting realistic goals and helps in comparing it with some benchmark.

This research is intended to make intelligent business decisions by following Business intelligence (BI) techniques and will reduce the dimensions of the raw dataset using deep learning approach. BI is a structure that changes the stock data into important information that improve the organization's essential basic leadership method [8]

Deep learning is a rising zone of Machine Learning (ML) research. It comprises multiple hidden layers of Artificial Neural Networks (ANNs). The deep learning framework applies nonlinear changes and model considerations of an anomalous state in generous databases. The ongoing headways in profound learning structures inside various fields have just given huge commitments to artificial intelligence[7]

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts [5]. Information distribution center needs to expand capacity after a period of time where multidimensional information are stored that is very expensive. If the dimensions of the data can be reduced, this can increase the storage capacity of a warehouse. This is why dimensionally data reduction is an important phase.

Data is growing at an uncommonly high rate. As data grows, the importance of organization to manage it and make it process-able grows as well. As the number of information increases, it needs to get the right data at the right place for the right things. As storage capability is important for data warehouse, multidimensional data will be taken and dimensions will be minimized without losing any information. In this way, it might expand the capacity limit of an information distribution center and decrease the expense of maintaining it. Meanwhile, each and every information are created by enormous data, the aim of the framework is to conquer the knowledge from the database and help to make intelligent business decisions. This may help to separate the useful data from the raw data to make better business decisions.

For this unique work of data science, we are proposing a framework merging deep learning concept along with BI technology that will help the business industries to take better business decisions by the managerial figures and will reduce the manual managerial efforts drastically as well.

This study is focused on producing a prediction of where to invest for better profits for the industries. We tended to reduce the dimension of raw data to see those data in a better way. Then the actual data then goes through many deep neural networks as input to find the patterns for the better-predicted result as output helping the managerial figures of the organizations to learn better business strategies from the traditional data that makes no sense. This is a work, for the sake of better

profits as toward the day's end all industries are wanted to enhance their businesses and focus on their financial growth.

2 Related Work

Business intelligence is today's tech priority. Business Intelligence is the process of going from raw data to legible information. BI solution helps to transform raw data into actionable information that helps support business decision making. This can help firms to develop new opportunities. By identifying these opportunities, and implementing an effective strategy, can provide a competitive market advantage and long-term stability [10]

As well as BI is termed as a set of numerical and technique models for examination used for extracting data and valuable information from raw data to utilize disordered basic leadership [9]. It is portrayed as frameworks that gather information and offering organized data from numerous sources by modifying the specified time to accumulate important business data and enable their efficient use in managerial decision making process [3], by permitting dynamic enterprise to information look, recovery, examination and clarification of the requirements of material selections [6]. The seminal element of business intelligence could be a broad class of technological applications and processes for gathering, storing, accessing and analyzing information to assist its users to make the best choices [11]

Different kinds of method like data mining or artificial neural network used for business prediction. For a sale forecast feed forward multilayer perception (MLP) networks is used with one hidden layer together with the back-propagation training method [4]. To get the future sale assumption, the previous data of some recent weeks inserted in the input layer. The next week sale prediction is the only result in the output layer. We applied linear regression here to predict sales and decision on investment.

Principal component analysis is perhaps the oldest and definitely the foremost widespread technique for computing lower-dimensional representations of variable information. This technique is linear within the sense that the elements are linear combos of the initial variables (features); however, nonlinearity within the information is preserved for effective visualization. The technique is conferred as associate degree reiterative computation of the direction of highest variation followed by projection onto the perpendicular hyper plane. This quickly provides some perpendicular directions that account for the bulk of the variation within the information, giving an occasional dimensional illustration of the information. A complete set of principal components is viewed as a rotation within the original variable area. In this paper, the concept of the geometric and dimension reduction properties of PCA is applied to the information to minimize the size and visualize it.

During the last decade, life sciences have undergone an incredible revolution with the accelerated development of high technologies and laboratory instrumentations. An honest example could be a medicine domain that has experienced a forceful advance since the arrival of complete ordering sequences [2].This post-genomics era has an effect on the event of recent high-throughput techniques that are generating huge amounts of information that has exponential growth of the many biological databases.

Linear Regression is an algorithm of machine learning followed by supervised learning. It performs a regression task. Regression models a target prediction data depending on independent variables. It's largely used for tracing out the connection between variables and prognostication. Regression differs depending on the nature of relationship between independent and dependent variables; they are considering and therefore the range of independent variables getting used. Linear regression performs an operation to predict a dependent variable value depending on a given independent variable.

One form of network sees the nodes as 'artificial neurons'. These are mentioned as Artificial Neural Networks (ANNs). An artificial neuron is additionally a method model affected among the natural neurons. Natural neurons receive signals through synapses settled on the dendrites or membrane of the neuron. Once the signals received are robust enough (surpass a specific threshold), the neuron is activated and emits a sign although the axon. This signal might be sent to a unique synapse and should activate different neurons. The complexity of real neurons is extraordinarily abstracted once modeling artificial neurons. These essentially contains inputs (like synapses), that are increased by weights (strength of the individual signals), then computed by a mathematical relation that determines the activation of the neuron. Another part computes the output of the artificial neuron. ANNs combine artificial neurons. The higher weight of an artificial neuron is, the stronger the input which is multiplied by it will be. Weights can also be negative, so we can say that the signal is inhibited by the negative weight. Depending on the weights, the computation of the neuron will be different. By adjusting the weights of an artificial neuron, we can obtain the output we want for specific inputs. However, when we have an ANN of hundreds or thousands of neurons, it would be quite complicated to find by hand all the necessary weights. Thus, we can find algorithms that can adjust the weights of the ANN in order to obtain the desired output from the network. This process of adjusting the weights is called learning or training [1].

Deep learning is an emerging area of Machine Learning (ML) research. It comprises multiple hidden layers of Artificial Neural Networks (ANNs). The deep learning methodology applies nonlinear transformations and model abstractions of high level in large databases [7]. Currently, Deep Learning is considered as an efficient business approach.

3 Method Development

The method of this research illustrates two main points: the way of collecting or generating data and the process of analyzing the data. We designed the following model to predict an intelligent business decision and to reduce the dimension of raw data that will reduce manual managerial efforts for better business decisions.

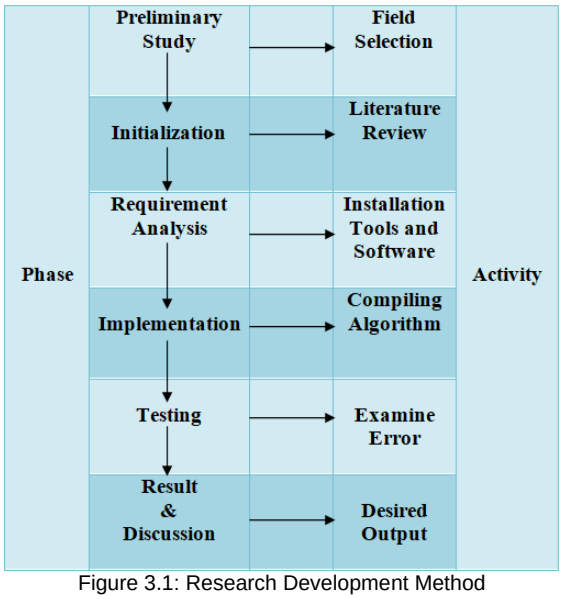

Figure 3.1: Research Development Method

PCA has been chosen for dimensionality reduction which also helps in data compression and hence reduced storage space. It reduces computation time. It also helps remove redundant features which create a linear orthogonal transformation of correlated data in one frame (coordinates system) to uncorrelated data in another frame. The huge dimensional data can be transformed and approximated with a few dimensions. PCA finds the directions of maximum variance in high-dimensional data and projects it onto a smaller dimensional subspace while retaining most of the original information. If the data is noisy, PCA reduces noise implicitly while projecting data along the principal components.

LR model is selected to predict intelligence decisions because Linear Regression is a well-supervised learning algorithm which is used to predictions problems. LR is very easy and intuitive to use and understand. It finds the target variable by finding the best suitable fit line between the independent and dependent variables. It is an extremely simple method that can map an N-dimensional signal to a 1-dimensional signal. It works well if the data has a clear linear trend. Even when it does not fit the data exactly, it can be used to find the nature of the relationship between the two variables.

3.1 Workflow of Our Research

To reduce the dimensions of the dataset, the following workflow is used based on Principal Component Analysis (PCA):

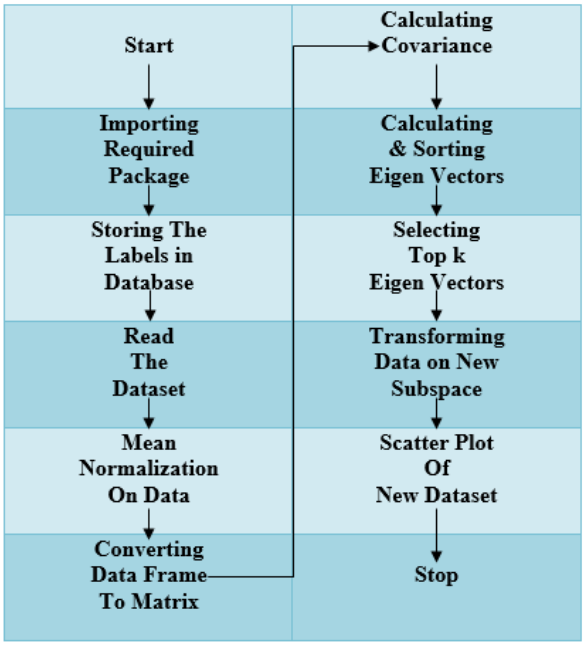

Figure 3.2: Workflow based on Principal Component Analysis (PCA) for Data Dimensionality Reduction.

The Linear Regression is used for predicting intelligence business decision. We used Linear Regression to predict decisions as the following workflow:

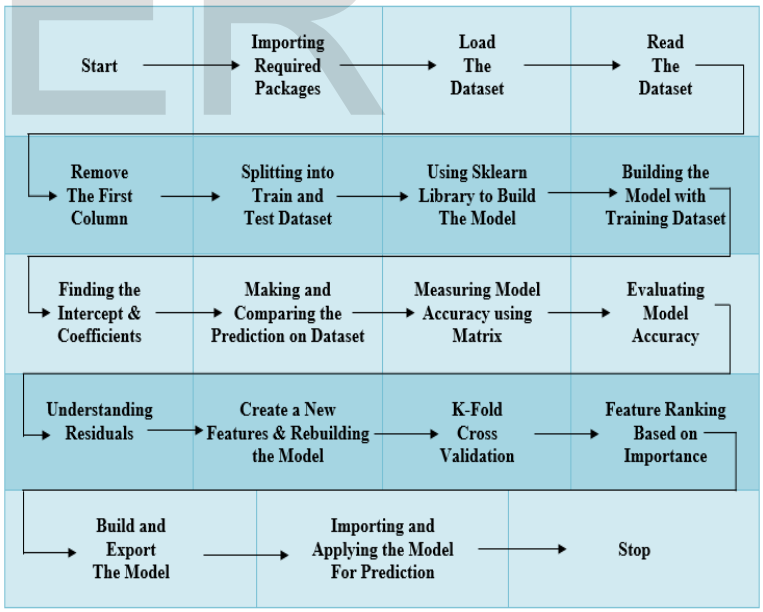

Figure 3.3: Workflow based on Linear Regression (LR) model for Intelligent Business Decisions.

.

4 Implementation and Analysis

The research and analysis refers to the overall strategy that we choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring we will effectively address the research problem; which constitutes the blueprint

for the collection, measurement and analysis of data.
We collected a dataset from the kaggle of a sales report and modified it later. Over six thousand cases of different transactions are included here. It consists of sales information of three products that are gathered from three different stations. Our calculation depends on the data to find out the best possible sales from the investment. We implemented this framework by using Python language at Jupyter Notebook (Anaconda Platform).

Research Implementation Procedure based on Algorithms

The research and analysis refers to the overall strategy that we choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring we will effectively address the research problem; which constitutes the blueprint for the collection, measurement and analysis of data.
We collected a dataset from the kaggle of a sales report and modified it later. Over six thousand cases of different transactions are included here. It consists of sales information of three products that are gathered from three different stations. Our calculation depends on the data to find out the best possible sales from the investment. We implemented this framework by using Python language at Jupyter Notebook (Anaconda Platform).

Research Implementation Procedure based on Algorithms

4.1 Data Dimensionality Reduction Implementation Procedure based on PCA Algorithm

Step I : Start
Step II : Import libraries
Step III : Read data
Step IV : Extract the last column of the dataset
Step V : Mean normalization –

$$x'=(x-\text{average}(x))/(\max_{fo}(x)-\min_{fo}(x))$$

Where, x is an original value and x′ is the normalized value.
Step VI : Convert data frame to matrix
Step VII : Calculate covariance –
S=(1/n)*XX^T
Where,
S is the variance-covariance matrix
[XX]^T is the deviation sums of squares and cross product matrix
n is the number of scores in each column
Step VIII: Calculan of the original matrix X.

Calculate Eigenvalues and Eigenvectors
Eigenvalue, T(V) = λ(V)
Where,
T is a linear transformation from a vector space
V is an eigenvector of T
λ is a scalar in the field known as the eigenvalue.
Eigenvector, A(V) = λ(V)
Where,
A is the square matrix
V is the column vector
λ is a scalar in the field known as the eigenvalue.
Step IX : Sort Eigenvalues in decreasing order
Step X : Select top K Eigenvectors; where, K = 2
Step XI : Transform data into new subspace
Step XII : End

4.2 Intelligence Business Prediction Implementation Procedure based on LR Model Algorithm

Step I : Start
Step II : Import Packages
Step III : Read Data
Step IV : Remove index column
Step V : Build a simple linear regression model –
Y=mX+b
Where,
m is the slope
b is the y-intercept
Y is the dependent variable
X is the explanatory variable.
Step VI : Split into train and test dataset
Step VII : Import Linear Regression from the sklearn linear model
Step VIII: Import train dataset
Step IX : Calculate intercepts and coefficients
Step X : Predict the test dataset

Step XI : Finding residuals
Residuals= Actual Value - Predicted Value
Step XII : Import matrixes from sklearn

Step XIII: Calculate RMSE –

$$RMSE=\sqrt{(1/n\sum_{(i=1)}^{n}[\![(f_i-o_i)]\!]^2)}$$

Where,
n is the number of samples
f is the forecasts
o is the observed values.
Step XIV : Calculate Root squared between test and predicted data
R-squared = Explained variation / Total variation
Step XV : Import libraries to plot residuals
Step XVI : K-Fold Cross Validation
Step XVII : Select features by using –

$$\_=[2]\_/((1-[2]\_))$$

Where,
F is the F-Test Value

is the number of regression parameters
Step XVIII: Build and export the model
Step XIX : Import and apply the model for prediction
Step XX : End

5 FINDINGS

Our data set consists of the sales information of three types of mineral water from three different branches of a super shop. Here we imported the Pandas data frame in the jupyter note-book. Pandas data frame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).

Reading the dataset and storing the labels in a label data frame

Pandas data frame consists of three principal components, the data, rows, and columns. Here is the 15 cases of our data set loaded dataset into Pandas data frame. We separated the labels given in the last column, for easy calculations on the data frame. Extracting the last column of the label and storing it into a label data frame for further use.

	Item1	Item2	Item3	Sales	Counter
0	230.1	37.8	69.2	22.1	Counter-1
1	44.5	39.3	45.1	10.4	Counter-2
2	17.2	45.9	69.3	9.3	Counter-3
3	151.5	41.3	58.5	18.5	Counter-1
4	180.8	10.8	58.4	12.9	Counter-1
5	8.7	48.9	75	7.2	Counter-2
6	57.5	32.8	23.5	11.8	Counter-3
7	120.2	19.6	11.6	13.2	Counter-2
8	8.6	2.1	1	4.8	Counter-1
9	199.8	2.6	21.2	10.6	Counter-2
10	66.1	5.8	24.2	8.6	Counter-1
11	214.7	24	4	17.4	Counter-3
12	23.8	35.1	65.9	9.2	Counter-2
13	97.5	7.6	7.2	9.7	Counter-1
14	204.1	32.9	46	19	Counter-2

The dataset after labeling:

Item1	Item2	Item3	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7
204.1	32.9	46	19

Figure 5.1: Label the data from the raw dataset for calculation

Transforming data on new subspace where the dataset is reduced dimensionally

We converted the data frame into matrix by performing mathematical operations. We then calculated the covariance, eigenvalues and eigenvectors of that featured matrix. After calculating eigenvalues and eigenvectors, we selected top k eigenvectors and sorted the eigenvalues in decreasing order. And the following is the reduced dataset of the raw data performed by PCA.

X	Y	label
84.2442	-40.1176	Counter-1
-102.218	-22.6466	Counter-2
-128.892	-48.2631	Counter-3
5.35206	-33.3506	Counter-1
34.1193	-21.377	Counter-1
-137.318	-54.7373	Counter-2
-89.7449	0.0778957	Counter-3
-27.4561	17.5054	Counter-2
-139.679	30.9033	Counter-1
52.0158	16.933	Counter-2
-81.4893	9.32721	Counter-1
66.9529	25.4818	Counter-3
-122.468	-41.1155	Counter-2
-50.484	25.3987	Counter-1
57.5334	-17.2692	Counter-2

Figure 5.2: After reducing the dimensions of data by performing PCA.

Scattering plot of the new dataset for data visualization

Scatter plot is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present. Here we converted the 4-dimensional space into 2-dimensional space (principal component 1 and principal component 2) of our dataset. In this diagram, we showed the graphical representation of reduced variances of our dataset. This scatter-plot helps to get a visualization of the entire dataset into a meaningful pictorial view.

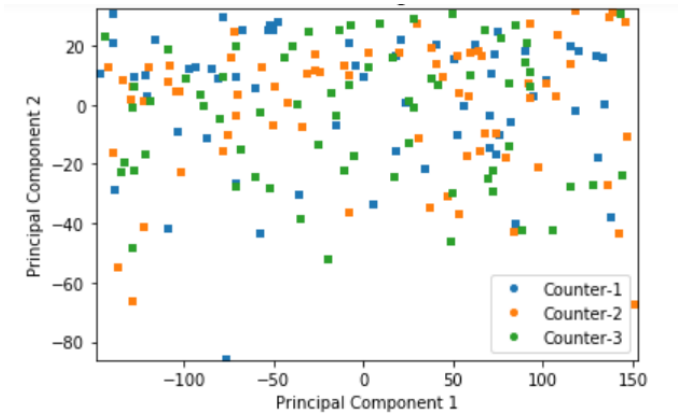

Figure 5.3: Data visualization after reducing dimensions by principal component analysis.

An intelligent decision has been produced for three counters after completing the linear regression. The prediction for investment of three items stands 472.204263, 195.721467 and 31.696670and the sales would increase into 293.603424.

Understanding Sales Plots

In order to visualize data from a Pandas Data Frame, we must extract each series and often concatenate them together into the right format. In this plot diagram, we showed the sales data from our data set.

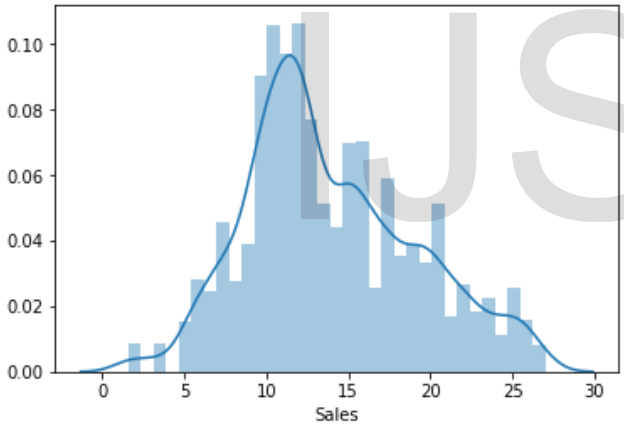

Figure 5.5: Sales plot of the data.

Understanding Residuals Plots

A residual is the difference between actual and predicted values. Residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in residual plots are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. In this diagram, we saw that our residual plots are also dispersed around the horizontal axis so that we can say that our linear regression model is appropriate for our data.

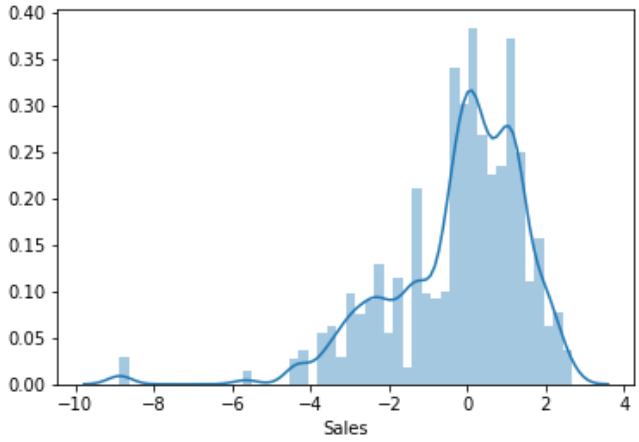

Figure 5.6: Residuals plot of the data

6 Conclusion

From the above research discussion, we see that how we can transform raw data into meaningful and useful information using business intelligence tools. Business Intelligence can be a great tool for reporting, benchmarking and doing data analysis for making sound business decisions. At senior managerial levels, it is used for making strategies and at lower managerial levels, it helps individuals to do their day-to-day job. According to Gartner, global revenue in the business intelligence and analytics software market was forecasted to reach $18.3 billion in 2017, an increase of 7.3 percent from 2016, according to the forecast from Gartner, Inc. By the end of 2020, the market is forecasted to grow to $22.8 billion. So we can see that analytics and business intelligence is the top technology priority for CIOs and demand for such data mining tool will continue to rise in the future. In conclusion, an attempt has been made to help the industry heads to take optimal decisions at the right time at the right place which will increase their business profits.

There are some good scopes in this paper where researchers can work in future and make it more updated and convenient. Researchers can use more complex and different kinds of dataset to get more accuracy whereas our dataset is rather more simple and smaller. They can use the other existing dimension reduction techniques and regression models for predicting the intelligent decisions. There is also a possibility of building a Graphical User Interface (GUI) based on the framework to form an entire intelligent business system that will help the users to easily manage and monitor their business growth.

Acknowledgment

We would also like to show our gratitude to the Authors for sharing their pearls of wisdom with us during the course of this research, and we thank “anonymous” reviewers for their so-called insights.

REFERENCES

[1] Carlos Gershenson, (2003), 'Artificial Neural Networks for Beginners'. Available from:https://arxiv.org/ftp/cs/papers/0308/0308031.pdf. [20 Aug 2003]

[2] C.O.S. Sorzano, J. Vargas & A. Pascual-Montano, 2014, "A survey of dimensionality reduction techniques", Natl. Centre for Biotechnology (CSIC); C/Darwin, 3. Campus Univ. Autónoma, 28049 Cantoblanco, Madrid, Spain; Available from-https://arxiv.org/ftp/arxiv/papers/1403/1403.2877.pdf

[3] Den Hamer, P. (2005). The organization of Business Intelligence. The Hague: SDU Publishers.

[4] Frank M Thiesing, Ulrich Middelberg, Oliver Vornberg, (1995), 'Short Term Prediction of Sales in Supermarket'. Department of MathematicsComputer Science University of Osnabruck. Available From: http://luna2.informatik.uni-osnabrueck.de/papers_pdf/icnn_95.pdf

[5] Iranmanesh, S.; Rodriguez-Villegas, E. (2017). "A 950 nW Analog-Based Data Reduction Chip for Wearable EEG Systems in Epilepsy". IEEE Journal of Solid-State Circuits. Available From: https://spiral.imperial.ac.uk:8443/handle/10044/1/48764

[6] Nofal, M., &Yusof, Z. (2013). Integration of Business Intelligence and Enterprise Resource Planning within Organizations. Technology, Vol. 11, pp. 658-665.

[7] R. Vargas, A. Mosavi, L. Ruiz, Deep Learning: A Review, Advances in Intelligent Systems and Computing, (2017). Available From:https://www.researchgate.net/publication/318447392_DEEP_LEARNING_A_REVIEW

[8] Singh, H., &Samalia, H. V. (2014). A Business Intelligence Perspective for Churn Management. Procedia-Social and Behavioral Sciences, 109, 51–56. doi:10.1016/j.sbspro.2013.12.420. Available_from:https://www.researchgate.net/publication/260007963_A_Business_Intelligence_Solberg Søilen, K. (2015). A place for Intelligence studies as a Scientific Discipline, Halmstad, Sweden. Journal of Intelligence Studies in Business, Vol. 5(3), pp. 35-46.

[9] Vercellis, C. (2013). Business Intelligence: Data mining and optimization for Decision Making. Amirkabir University Press, 2nd Edition.

[10] Vikas Kumar, 2013, San Diego State University, California. Available From: http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/4212/Kumar_Vikas.pdf;sequence=1

[11] Wixom, B. and Watson, H. 2010. The BI-Based Organization. International Journal of Business Intelligence Research, Vol. 1(1), pp. 12-24.