Data Completeness Prediction by Deep Learning

Jaouad MAQBOUL[1, *], Bouchaib BOUNABAT[2, *]

[*]AL QualSADI Team-ENSIAS, Rabat IT Center, Mohammed V University, Rabat, Morocco

Abstract.

This article describes an approach to identify the tangible and intangible im-pact of better data quality, in an enterprise architecture context without for-getting the cost resulting from the improvement of this data. the goal is to measure the impact and cost of improving business processes, quantitatively, to help decision-makers make good decisions and carry out their strategy, this approach will facilitate the choice of candidate quality projects to be ex-ecuted by minimize cost of improvement, an JEE java web application is de-veloped to meet our need.

Keywords: Data quality assessment and improvement, artificial neural network, deep learning Introduction.

1. Introduction

Formerly we speak of budgetary capital and the money invested, according to IBM's the expense of poor-quality data is enormous, and estimate to $3.1 trillion per year for the US economy, $611bn per year is lost in the US in poorly targeted mailings and staff overheads alone [1], nowadays we talk about of data, information, knowledge to wisdom capital, the immensity of the data produced by the compa-ny during its existence is an essential heritage of which it cannot do without it, 90% of data produced so far, in the last two years [2], the global big data analytics market was valued at US $ 37.34 billion in 2018 and is expected to reach US $ 105.08 billion by 2027 and 58% of companies adopt Big Data Analytics [30]; by the way Big Data is characterized by properties which are (Variety, Variabil-ity, Complexity, Value) [31], by these properties there will surely be a lot of erro-neous, incomplete, inaccurate data and it is the duty of compagnies to correct them as the cost of poor quality data is enormous and could damage image of company with customers and partners; for this purpose that practitioners and scientists, identify quality as a strategic element that will create added value, and satisfy the client to humanity, in meeting their needs, and gaining here trust for making a financial gain [3].

The organization of this paper is addressed as follows: section I a comparative table of research work on the impact and cost of quality, Section II, presents data quality definitions and dimensions; Section III the contribution and the complexi-ty of the completeness improvement; Sections IV describes our approach of using the artificial neural network on completeness prediction, the conclusions and future work are summarized.

2. Related Work

Table 1. Related work

Works	Data	Process	Cost		Impact	
			Poor data	Improvement	Financial	non-Financial
Data Quality for the Information Age (Redman, 1996)	√		√			
Enterprise Knowledge Management: The Data Quality Approach (Loshin, 2001)	√		√			
Towards Quantifying Data Quality Costs (Kim & Choi, 2003)	√		√	√		
A Classification and Analysis of Data Quality Costs (Eppler & Helfert, 2004)	√		√	√		
Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) (McGilvray, 2008)	√		√		√	
Information Quality Applied (English, 2009)	√		√			
DAMA Book of Knowledge (DAMA, 2009)	√		√		√	
Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits (English, 2011)	√	√	√			
The Costs of Poor Data Quality (Haug, 2011)	√		√			
Understanding the Financial Value of Data Quality Improvement (Knowledge Integrity, 2011)	√		√		√	
Measuring the Business Value of Data Quality (Gartner, 2011)	√				√	
How to Monetize, Manage, and Measure Information as an Asset for Competitive X. (Laney, 2017)	√				√	

The comparison criteria being the aspects dealt with, namely: the costs of poor quality the impact of improved data quality and the costs of improving quality. This for the processes and its data.

It can be concluded that the literature in the field of impact assessment and efficiency analysis of a preliminary program / project to improve data quality is broken down into the following aspects:

1. The negative value of the data, namely the costs associated with poor quality;
2. The financial / business value of improving the quality of the data;
3. The costs associated with improving the quality of the data.

However, this work does not offer metrics or aggregated indicators that allow us to analyze these aspects, to lead, to choices of implementation of initiatives to improve the quality of the data, the work done in this direction was based on the economic side, which makes these models difficult to apply and adapt to a range of contexts.

The aggregation of the different impact and cost factors is required in order to provide a synthesized view and thus simplify the decision-making process in relation to the quality improvement initiatives to be implemented.

3. Data Quality Dimensions Classification, Study Case Data Completeness

3.1 Data Quality

We have chosen some definitions of quality,quality is the degree to which a set of inherent characteristics fulfills requirements [4]. Philip Crosby says: Conformance of requirements of the customer [5].W. Edwards Deming: A predictable degree of uniformity and dependability, at low cost and suited to the market [6]. Joseph Juran: fitness for use [7].Referred to [8], quality is “the extent to which information still meets the requirements and expectations of customers [9, 10].

3.2 Classification of the Quality Dimension

The researchers made a classification from three to four categories of data quality by Wang and Strong: intrinsic, contextual, representational and accessibility [11], in this study we choose completeness from contextual.

1. Intrinsic: Denotes that data is linked to the quality that the data has on its own. This aspect of quality is independent of the user’s perspective and context.
2. Contextual: Points to the requirement that data quality must be considered within the context of the task at hand.
3. Representational: includes aspects related to the format of the data, and meaning of data [36], in such a way that they are interprétable, easy to understand.
4. Accessibility: emphasizes the important role of systems; that is, the data must be accessible as well as the system in addition to its security.

3.3 Chosen dimension, Completeness

This study will survey on one of critical dimensions; accuracy, consistency, completeness, and timeliness, which are considered as a fundamental dimension and improvement process in information systems [36], data completeness can block the business process than other critical dimensions.

Among the most common definitions of completeness is the extent to which data is not missing and is of sufficient breadth and depth for the task at hand [11], it means all the requisite information available.

In the next session, we discuss the impact of completeness improvement and how will cost to the company; but the cost will be difficult to calculate and estimate, so we choose the complexity of completeness improvement; it is already proven that the cost of improvement has a linear relationship with complexity [12, 13],and the improvement will target the data managed by these processes.

In the next section we extracted the questions in our survey to see the impact of improving the data completeness on the company in terms of costs and gains.

4. The Complexity And Contribution Of Improving Completeness

4.1 Impact of improving data completeness

Good-quality data has several beneficial impacts on organizations in Decision making, Productivity, Compliance and Marketing [35], here we cite the questions to answer the impact of data completeness,is improving data completeness

1. Will reduce the cost?

Good data quality cost to company money and time, the cost can be divided into cost of low data and the cost of improving data [14]; the second class divided into prevention costs, Detection costs, repair costs; by default, good data, minimize these costs and we don't waste time and resources to deal with poor data.

2. Will increase profitability and efficiency?

A report made by General Accounting Office 1991, companies that used total quality management practices, improved profitability [15]; so good data in-crease profitability. When collaborators are committed in a work environment; in which teamwork is emphasized, and where quality are the goal; the work will be fluid and smoother than the one in which quality comes after the incident.

3. Will enhanced customer satisfaction?

In order to survive, the company must continuously improve its ability to serve their customers, and have the advantage of competing better than existing competitors [16]. good quality data keeps customers loyal without investing in advertising or any other method of influence, even if the staff leave the compa-ny; the services is delivered in time no matter the employers behind the process

4. Will improve transverse processes (Cross-Functional Process)?

The processes that are the true value of the business are precisely those that cover multiple organizational domains, applications, and IT systems. Most of the time, processes are divided into interdepartmental because once a process is transferred to the next service; it is difficult to track and manage, so the im-provement of the data managed by these processes will improve the daily opera-tions for interdepartmental.

5. Will facilitates decision making and strategy change?

Today, most companies use data to make decisions about their business. The reason they're leading the way is that they've gained a strategic advantage over their rivals simply by shifting their focus to data rather than relying on business acumen alone. Good data make it easy for analysts provide good information to decision-makers to make good decisions and even make changes to their strate-gy in the short or longterm.

6. Will has a direct impact over time?

It depends on the context of the entity, quality data, in our context, complete-ness, will have an impact in the short-term or in the long-term.For example, accumulated market data can save the company large sums, in the short-term or, on the contrary, mark its death in front of its competitors; while a strategy based on complete data will have a long-term impact.

7. Will increase security compliance?

Almost all privates, data protection and compliance programs are data-based. And the quality of this data is important. If it is low, you are exposed to risks, but the contrary will help to reinforce security, generate savings by refining processes, reduce compliance time by streamlining the reporting process and Reduce the response time.

4.2 The complexity for data completeness improvement

the cost of a data quality comes from the cost of prevention, direct cost and cost of improvement and others, for our study we will try to ask questions about the cost of improvement of the completeness, how cost to improving data completeness by:

1. Finding of source of data to complement or contradict the data?
   The companies are looking for to complete their data; there is some data that can be completed, by older backup after failed migration, or knowledge bases that are had been capitalized after years of experience including algorithms and instructions for employees, to deliver faster service, and increasing satisfaction of customer; the fact that experts contribute to this knowledge base, doesn’t improve that their knowledge accumulated are archived, documented and shared with all employers; further their knowledge can contradict the archived knowledge and correct it.
2. The nature of data?
   Most companies rely on huge amounts of data; about their suppliers, employees, process data, videos, audio, PDFs, knowledge and more; business processes handle several types of data ranging from golden data, master data to unstructured data.The complexity increases from master data that manipulate some tables like employees, customers, products or suppliers; these tables are key business entities that support transactions in operational systems [17], to the most expensive, unstructured data; these tables contains information that does not have a predefined data model or is not organized in a predefined

manner; that's the fact because there is data entered manually, by machine directly or after back-end processing based on a faulty algorithm in a knowledge base.

3. How the process is executed?

   Business Process Automation (BPA) is assumed to enhance organizational efficiency by decreasing levels of effort and elimination of redundant processes and procedures [18]. The complexity varies according to the way the process is executed manually or automatically.

4. Identifing attributes of data that have great weight?

   Identify attributes that are consistent across different data sources, reduces the complexity of Data matching [26], data cleansing [27], Data profiling [28], Data deduplication [29].

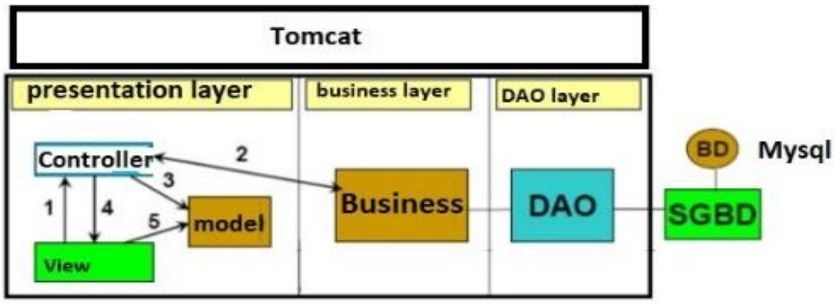
In the next session, we will discuss about deep learning to our neural network in order to classify the prediction of completeness and calculated after by regression.

5. Proposed Architecture

5.1 Technical architecture

Architecture is divided into 3 layers: the presentation layer which provides the interface with the user, the business layer which encloses the business processes and the DAO layer, through a connector, is responsible for the operations of managing records at the level database. this layered architecture allows for easy maintainability and scalability [32].

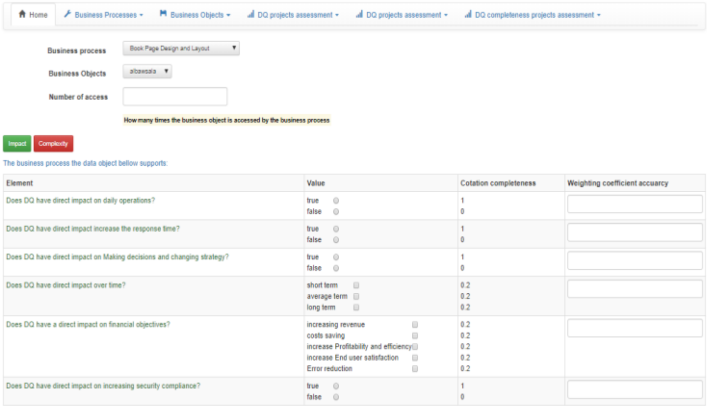
Figure 1. technical architecture of the jee application



5.2 implementation phase

we present in this section the screenshots of the platform

Figure 2. assessment of the positive impact factor



Business process: Book Page Design and Layout

Business Objects: albawsala

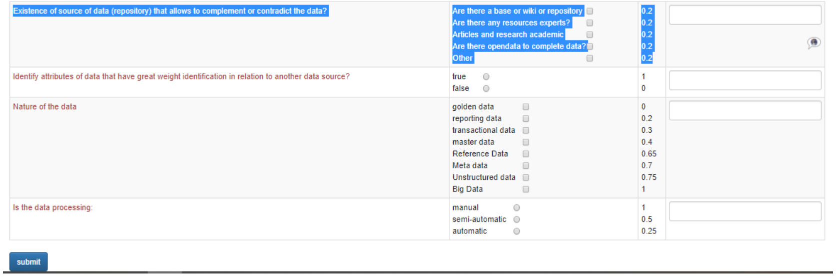
Number of access:

How many times the business object is accessed by the business process

The business process the data object bellow supports:

Element	Value	Cotation completeness	Weighting coefficient accuarcy
Does DQ have direct impact on daily operations?	true ○ false ○	1 0	
Does DQ have direct impact increase the response time?	true ○ false ○	1 0	
Does DQ have direct impact on Making decisions and changing strategy?	true ○ false ○	1 0	
Does DQ have direct impact over time?	short term □ average term □ long term □	0.2 0.2 0.2	
Does DQ have a direct impact on financial objectives?	increasing revenue □ costs saving □ increase Profitability and efficiency □ increase End user satisfaction □ Error reduction □	0.2 0.2 0.2 0.2 0.2	
Does DQ have direct impact on increasing security compliance?	true ○ false ○	1 0	

Figure 3. assessment of the complexity of implementation



Existence of source of data (repository) that allows to complement or contradict the data?	Are there a base or wiki or repository □ Are there any resources experts? □ Articles and research academic □ Are there opendata to complete data? □ Other □	0.2 0.2 0.2 0.2 0.2	
Identify attributes of data that have great weight identification in relation to another data source?	true ○ false ○	1 0	
Nature of the data	golden data □ reporting data □ transactional data □ master data □ Reference Data □ Meta data □ Unstructured data □ Big Data □	0 0.2 0.3 0.4 0.65 0.7 0.75 1	
Is the data processing:	manual ○ semi-automatic ○ automatic ○	1 0.5 0.25	

6. Artificial Neural Networks

6.1 Principe of ANN - neural network

Deep learning, and machine Learning they are new, and effective strategies commonly to maximize profits in the market; neural networks by nature are effective in finding the relationships between data and using it to predict (or classify) data [19]. An artificial neural network (ANN) involves numerous processors operating in parallel arranged in levels [21]. The first level receives the input information. Each successive level receives the output of the preceding level, the last level produces the output of the system which will be the prediction.

6.2 The data completeness prediction by the ANN

ANN uses multiple layers that approximate complex mathematical functions to process data. A backpropagation neural network (BPNN) is a widely used learning algorithm in an ANN application that uses back-propagation of the error gradient to minimize the error of nonlinear functions of high complexity [25]. In this study, a BPNN algorithm has been adopted for predicting completeness after an improvement of it, which will have a favorable impact on the company at a cost (complexity), based on an experience in the company on the previous improvements.

The BPNN consists of three layers: an input layer, one or multiple hidden layers and an output layer; each layer contains a number of neurons, each of which receives inputs from neurons in the previous layer, or external inputs, and converts these into either an output signal [22, 24].

In our study the factors of impact and complexity are going to be the input of our network, in Figure 2 and Figure 3, we had used weight for each answer; the strong point for this propose came from the contextual fly of completeness; our proposal to let the neuron network learn from data to adapt these weight in order to minimize the cost in Figure 3 and approaching the completeness.

Backpropagation: Backpropagation is a method to adjust connection weights, to compensate for each error found during learning; the goal of backpropagation is to improve the prediction in most learning networks; so, error is calculated as the difference between the actual output and the predicted output [23, 34]. To decrease the error, move the weight in the opposite direction of the gradient, means the partial derivative of the J(w) with respect to weight w.

Forward propagation: Refers to, calculating and storing intermediate variables (including outputs) for the neural network within the models; in the order of the input layer to the output layer, we can choose a function activation for each layer; we use function activation to learn nonlinear complex functional mappings between the factors (inputs) and the predicted completeness (target outputs of our data.

7. Completeness Prediction By Neural Network

7.1 Learning from answers of the survey

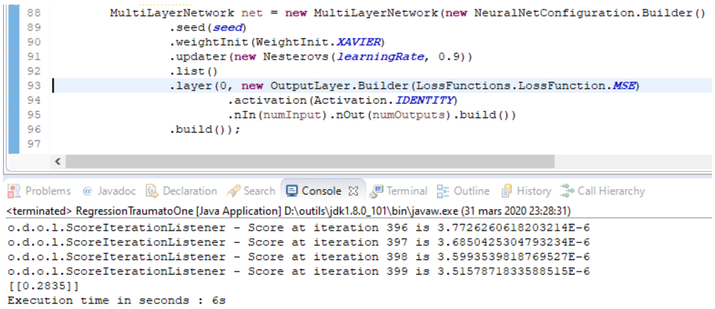
We have developed a survey to question the practitioners and responsables of processus of data quality, our survey is about the cost to improve data completeness of proceses and it’s impact on the company; we pushed at level to accumulate information about process in company; so the field company is added in order to have a learning for just this company; the type of company to generalize this learning for same company, the name of process, type of process (critical one, important, normal) and the department charged of this process; in order to generalize to other company, or same department, and other fields existed in our last survey to identify the factors of (impact, complexity) completeness improvement [20].

7.2 case study

7.1.1 Case 1.

In first, we have used one hidden layer with 10 neurons in order to predict the completeness; we have used the dl4j library, the value of predict completeness 0.2835 is more than the desired value, while the value calculate after improving is 0.25.

Figure 4. The predicted value of completeness for (1) layer



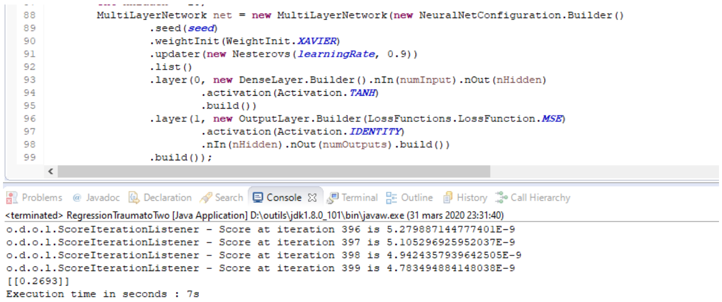
```
        MultiLayerNetwork net = new MultiLayerNetwork(new NeuralNetConfiguration.Builder()
                .seed(seed)
                .weightInit(WeightInit.XAVIER)
                .updater(new Nesterovs(learningRate, 0.9))
                .list()
                .layer(0, new OutputLayer.Builder(LossFunctions.LossFunction.MSE)
                        .activation(Activation.IDENTITY)
                        .nIn(numInput).nOut(numOutputs).build())
                .build());
```

```
<terminated> RegressionTraumatoOne [Java Application] D:\outils\jdk1.8.0_101\bin\javaw.exe (31 mars 2020 23:28:31)
o.d.o.l.ScoreIterationListener - Score at iteration 396 is 3.7726260618203214E-6
o.d.o.l.ScoreIterationListener - Score at iteration 397 is 3.6850425304793234E-6
o.d.o.l.ScoreIterationListener - Score at iteration 398 is 3.5993539818769527E-6
o.d.o.l.ScoreIterationListener - Score at iteration 399 is 3.5157871833588515E-6
[[0.2835]]
Execution time in seconds : 6s
```

7.1.2 Case 2.

After we decide to go with two layers, each layer with 10 neurons, the value of predict completeness 0.2693 is more than the desired value, while the value calculate after improving is 0.25.

Figure 5. The predicted value of completeness for (2) layer



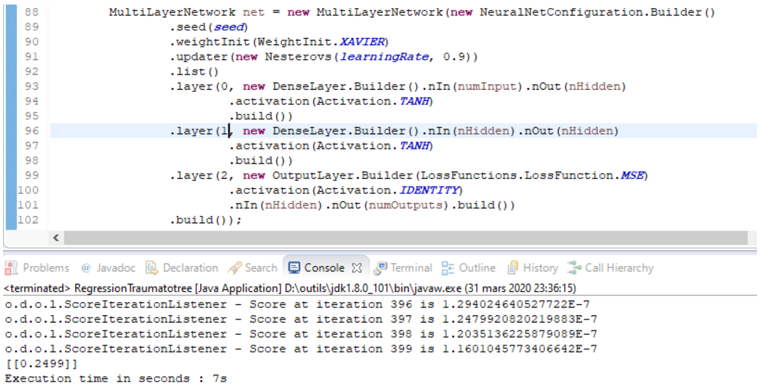
```
        MultiLayerNetwork net = new MultiLayerNetwork(new NeuralNetConfiguration.Builder()
                .seed(seed)
                .weightInit(WeightInit.XAVIER)
                .updater(new Nesterovs(learningRate, 0.9))
                .list()
                .layer(0, new DenseLayer.Builder().nIn(numInput).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(1, new OutputLayer.Builder(LossFunctions.LossFunction.MSE)
                        .activation(Activation.IDENTITY)
                        .nIn(nHidden).nOut(numOutputs).build())
                .build());
```

```
<terminated> RegressionTraumatoTwo [Java Application] D:\outils\jdk1.8.0_101\bin\javaw.exe (31 mars 2020 23:31:40)
o.d.o.l.ScoreIterationListener - Score at iteration 396 is 5.279887144777401E-9
o.d.o.l.ScoreIterationListener - Score at iteration 397 is 5.105296925952037E-9
o.d.o.l.ScoreIterationListener - Score at iteration 398 is 4.9424357939642505E-9
o.d.o.l.ScoreIterationListener - Score at iteration 399 is 4.783494884148038E-9
[[0.2693]]
Execution time in seconds : 7s
```

7.1.3 Case 3.

Add one hidden layer, so we had 3 hidden layers, the predicted completeness approximate the desired value 0.2499

Figure 6. The predicted value of completeness for (3) layer



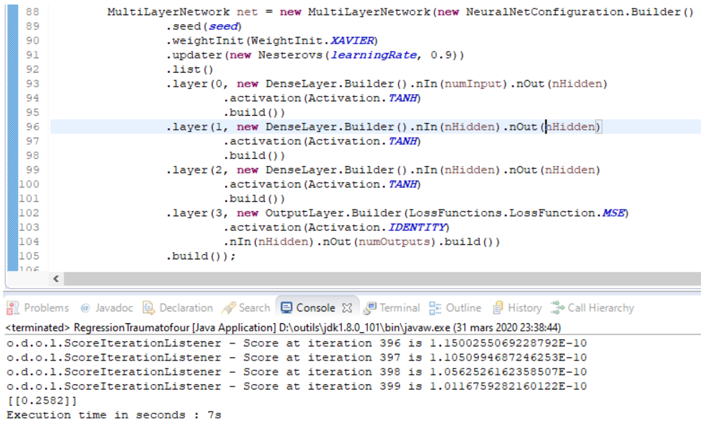
```
        MultiLayerNetwork net = new MultiLayerNetwork(new NeuralNetConfiguration.Builder()
                .seed(seed)
                .weightInit(WeightInit.XAVIER)
                .updater(new Nesterovs(learningRate, 0.9))
                .list()
                .layer(0, new DenseLayer.Builder().nIn(numInput).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(1, new DenseLayer.Builder().nIn(nHidden).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(2, new OutputLayer.Builder(LossFunctions.LossFunction.MSE)
                        .activation(Activation.IDENTITY)
                        .nIn(nHidden).nOut(numOutputs).build())
                .build());
```

```
<terminated> RegressionTraumatotree [Java Application] D:\outils\jdk1.8.0_101\bin\javaw.exe (31 mars 2020 23:36:15)
o.d.o.l.ScoreIterationListener - Score at iteration 396 is 1.294024640527722E-7
o.d.o.l.ScoreIterationListener - Score at iteration 397 is 1.2479920820219883E-7
o.d.o.l.ScoreIterationListener - Score at iteration 398 is 1.2035136225879089E-7
o.d.o.l.ScoreIterationListener - Score at iteration 399 is 1.1601045773406642E-7
[[0.2499]]
Execution time in seconds : 7s
```

7.1.4 Case 4.

We decide to add another hidden layer to identify the exact numbers of hidden layers, so know we had 4 hidden layers, the predicted completeness is 0.2583.

Figure 7. The predicted value of completeness for (4) layer



```
        MultiLayerNetwork net = new MultiLayerNetwork(new NeuralNetConfiguration.Builder()
                .seed(seed)
                .weightInit(WeightInit.XAVIER)
                .updater(new Nesterovs(learningRate, 0.9))
                .list()
                .layer(0, new DenseLayer.Builder().nIn(numInput).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(1, new DenseLayer.Builder().nIn(nHidden).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(2, new DenseLayer.Builder().nIn(nHidden).nOut(nHidden)
                        .activation(Activation.TANH)
                        .build())
                .layer(3, new OutputLayer.Builder(LossFunctions.LossFunction.MSE)
                        .activation(Activation.IDENTITY)
                        .nIn(nHidden).nOut(numOutputs).build())
                .build());
```

```
<terminated> RegressionTraumatofour [Java Application] D:\outils\jdk1.8.0_101\bin\javaw.exe (31 mars 2020 23:38:44)
o.d.o.l.ScoreIterationListener - Score at iteration 396 is 1.1500255069228792E-10
o.d.o.l.ScoreIterationListener - Score at iteration 397 is 1.1050994687246253E-10
o.d.o.l.ScoreIterationListener - Score at iteration 398 is 1.0562526162358507E-10
o.d.o.l.ScoreIterationListener - Score at iteration 399 is 1.0116759282160122E-10
[[0.2582]]
Execution time in seconds : 7s
```

The experience has shown that three hidden layers are the most suitable for our problematic, and we had the best prediction we can have.

8. RESULTS AND DISCUSSION

Our survey is a tool for capitalizing the cost and impact of improving the completeness of business processes in the company, we can have several costs and impacts for the same business process over time, we can group them by business services or department or services, from this database proven after improvement of completeness, the neuron network will use it as input to predict a value, in our study will be the value of completeness based on the factors of the cost and impact, we used multiple hidden layer levels and case 3 (hidden three layer neural network) accurately predicted the completeness value.

9. Conclusion And Future Work

Our approach is based on the impact and complexity of data quality improvement, in our case completeness; based on neural networks as factors are numerous and each factor influences it company; then each company will build its own model and can be generalized on another company or department.

After each improvement, it's necessary to compare predicted value, with the real one; update the value, and recalculate if the predicted will the completeness and have a precision of completeness and prediction over time.

Some future work: (i) the generalization of our approach for more dimensions of data quality (ii) application of our approach by dimension category (iii) combine the shapley value with our approach to give decision-makers to choose the data processes or departments or departments to improve and expand it to the dimensions of quality.

References (TNR 14pt., bold)

[1] Anders Haug, Frederik Zachariassen, Dennis van Liempd . (2010), The costs of poor data quality.

[2] Ibrar Yaqoo, Ibrahim Abaker Targio Hashem, Abdullah Gani, Salimah Mokhtar, Ejaz Ahmeda,Nor Badrul Anuar, Athanasios V. Vasilakos. (2016), Big data: From beginning to future.

[3] Roland Kantsperger, Werner H. Kunz. (2014), Consumer Trust in Service Companies:a Multiple Mediating Analysis

[4] International Organization for Standardization, Quality Management Systems—Fundamentals and Vocabulary (Geneva: ISO Press, 2005), in Project Management Institute, Inc., A Guide to the Project Management Body of Knowledge (PMBOK Guide), 4th ed.

[5] Square. (2008), PA: Project Management Institute, Inc., 190

[6] Crosby, P.B. (1979), Quality is free : The Art of Making Quality Certain, p 39.

[7] Deming, E.W. (1982), Quality, Productivity and Competitive Position. MIT Press, Cambridge.

[8] Juran, J.M. (1998), Juran on Leadership For Quality, p 15.

[9] Jaouad Maqboul, Bouchaib Bounabat. (2017), Towards a Completeness Prediction Based on the Complexity and Impact.

[10] Joseph M. Juran,A. Blanton Godfrey, JURAN'S QUALITY HANDBOOK (1979), p 21.

[11] Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers.

[12] Pipino, L.L., Lee, Y.W., Wang, R.Y. (2002), Data quality assessment. Commun. ACM 45(4), 211– 218.

[13] S. Arestie, R. Harris, N. Ricci, M.A. Schaffner, M.M. Shaw, C. Whitlock, D. Miller. (2013), A bottom-up modeling approach for the profit analysis of cellularized spacecraft architectures.

[14] Hvenegaard, A., Arendt, J.N., Street, A., Gyrd-Hansen, D. (2011), Exploring the relationship between costs and quality—does the joint evaluation of costs and quality alter the ranking of Danish hospital departments? Eur. J. Health Econ. 12(6), 541–551 .

[15] Anders Haug, Frederik Zachariassen, Dennis van Liempd, The costs of poor data quality,2011.

[16] Juran, J.M. (1998), Juran on Leadership for Quality, p 231.

[17] L Theresia, and R Bangun. (2017), Service quality that improves customer satisfaction in a university: a case study in Institut Teknologi Indonesi.

[18] Stephan E. Zoder, MIPP, Improving Enterprise Master Data Quality What Is The ROI?

[19] Sarfraz Ahmad Sirohey, Ahmed Imran Hunjra, Babar Khalid, Impact of Business Process

[20] Automation on Employees' Efficiency. (2012).

[21] D. Fister, j.c. mun, v. jagri c, t. jagri c .(2019), deep learning for stock market trading: a superior trading strategy.

[22] Tugce Karatas, Amir Oskoui, Ali Hirsa. (2019), Supervised Deep Neural Networks (DNNs) forPricing/Calibration of Vanilla/Exotic Options Under VariousDifferent Processes.

[23] Md. Haidar Sharif, Osman Gursoy. (2018), Parallel Computing for Artificial Neural Network Training using Java Native Socket Programming,.

[24] Massimo Buscema. (1998), Back Propagation Neural Networks, in Substance Use & Misuse,.

[25] Sang-Hoon Oh. (1997), Improving the Error Backpropagation Algorithm with a Modified Error Function.

[26] Alexander Semenov, Vladimir Boginski, Eduardo L. Pasiliao. (2019), Neural Networks with Multidimensional Cross-Entropy Loss Functions.

[27] J V N Lakshmi,. (2016) Stochastic Gradient Descent using Linear Regression with Python,.

[28] Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection (2012).

[29] Otmane Azeroual, Gunter Saake, Mohammad Abuosba. (2017), Data Quality Measures and Data Cleansing for Research Information Sys-tem.

[30] Wei Dai1, Isaac Wardlaw, Yu Cui, Kashif Mehdi, Yanyan Li, Jun Long, Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking.

[31] Shobha,S. C. Jain. (2018), Efficient Data Deduplication for Big Data Storage Systems.

[32] Gabriele Piantadosi, Stefano Marrone, Mario Sansone & Carlo Sansone. (2015), A secure, scalable and versatile multi-layer client–server architecture for remote intelligent data processing,

[33] Avita Katal Mohammad, Wazid, R H Goudar. (2017), Big Data: Issues, Challenges, Tools and Good Practices.

[34] D. E. rumelhart, G. E. hinton, and R. J. Williams. (2014), Learning Internal Representations by Error Propagation, 1989. Rumki Majumdar, Business decision making, production technology and process efficiency.

[35] Payam Hassany Shariat Panahy, Fatimah Sidi, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim and Aida Mustaph, A Framework to Construct Data Quality Dimensions Relationships, Indian Journal of Science and Technology.