Creating a Business Value while Transforming Data Assets using Machine Learning

Ivana Dimitrovska, Toni Malinovski

Faculty of Information and Communication Technology, FON University Skopje, Republic of Macedonia
ivana.dimitrovska@fon.mk, toni.malinovski@fon.mk

ABSTRACT

Machine learning enables computers to learn from large amounts of data without specific programming. Besides its commercial application, companies are starting to recognize machine learning importance and possibilities in order to transform their data assets into business value. This study explores integration of machine learning into business core processes, while enabling predictive analytics that can increase business values and provide competitive advantage. It proposes machine learning algorithm based on regression analysis for a business solution in large enterprise company in Macedonia, while predicting real-value outcome from a given array of business inputs. The results show that most of the machine learning predictive values for the desired process output deviated from 0 to 15% of actual employees' decision. Hence, it verifies the appropriateness of the chosen approach, with predictive accuracy that can be meaningful in practice. As a machine learning case study in business context, it contains valuable information that can help companies understand the significance of machine learning for enterprise computing. It also points out some potential pitfalls of machine learning misuse.

Keywords: Machine Learning, Business Value, Business Solution, Predictive Analytics, Regression Algorithm.

1. INTRODUCTION

Ever since the earliest beginnings of technological development, scientists have recognized the potential of machines for advanced computing and processing. Having in mind that learning that leads to reasoning is an essential ability associated with intelligence, machine learning (ML) has received much attention during the short history of computer science [1], [16]. ML enables computers to learn from large amounts of data without being explicitly programmed to do so [3], [11]. Hence, evidence already show its contribution to new intelligent solutions, such as self-driving cars [8], different information technology (IT) applications, from email-spam filters [13] to large-scale data processing [2], as well as solutions that solve important problems faced by the astronomical community [9], medicine [6], [24], etc.

On the other hand, academics wishing to understanding the nature of human learning are faced with complex topic that spans a spectrum of disciplines [4]. Human learning can be naturally divided in two categories: unconscious and intentional learning. When doing it intentionally, people have different approaches and methods of learning new things. For some of them learning can be acquired only by explicit processes as reading, others learn while listening or even through a figurative presentations. Still, all of them strive towards the same goal and eventually gain the desired results: to learn something they did not know before. Similarly, we can train machines to obtain new knowledge in different ways beyond standard algorithms [1], [3], [7].

The recent developments of ML techniques led to a widespread of applications in different areas, which can help employ a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible. Business applications are not excluded from this trend, while many companies try to use ML to manipulate and analyze their data assets to gain certain business values [10], [19]. Even more, the significance of ML application in business processes is very important and should be considered very carefully. The business process is single or sequence of tasks executed in a predefined order with predefined dependences, which produce a specific service [5]. Managing such process flows can be difficult and exhausting, which often cannot be handled by a single person [22]. Business processes improvement is a first step towards automation and one of the ways for increasing company’s market position. ML in business applications is an ultimate process improvement technique that can enhance services and products [14]. Everything that humans cannot achieve because of limited processing capabilities or biased approach while analyzing things, can be surpassed by intelligent processing and reasoning. However, one should always be aware of the potential challenges and problems while using ML in business applications.

The goal of this study is to emphasize benefits of ML usage in business context, as well as to point out some potential pitfalls of its misuse. Before one can plan to introduce ML into business solutions, there are several aspects that must be considered. Although ML provides an innovative approach, not every single business challenge should be tackled with ML techniques. On the other hand, when appropriate, resolving problems with the help of ML algorithms may definitely increase business values and provide competitive advantage for the companies. Therefore, this study tries to open up several new avenues for research in addition to recent studies [1], [2], [9], [14], [19], [25] while the proposed methods and chosen algorithm are verified via ML implementation into standard business solution for large enterprise company in Macedonia. Hence, the whole approach strives to create a business value during transformation of data assets via ML, practically implemented in the chosen company and its core business processes. This paper provides encouraging results, which open up a possibility of coexistence between standard and ML coding techniques into a single business solution, as well as positive feedback and end-user benefits.

2. THEORETICAL BACKGROUND

Modern datasets in business processes are rapidly growing in size and complexity, and there is a pressing need to develop solutions to harness this wealth of data and create additional business value for the organizations. Business intelligence and analytics may be used for processing of large and complex datasets [10], but ML can devise complex models and algorithms that can provide predictive analytics and uncover hidden insights [12], [16], [24], through learning from historical relationships and trends in the data [1], [7]. Hence, information mining supports business intelligence tools to transform information into knowledge [15]. It searches for interesting patterns and important regularities in large bodies of information [15]. Behind every logical information mining process stands a data mining algorithm responsible for the ML task execution.

As indicated in [16], [23], one of the basic ML tasks is classification - a method for mapping examples into predefined groups or classes. This task is one of the most frequently used tasks, especially related to business applications. It is often referred as supervised learning, because the classes are determined before the data examination process [17]. In [16], [20], [21] researchers elaborate additional ML domains, such as unsupervised learning and reinforced learning, but they all indicate that supervised learning is most widely used ML method [20], [21]. One of the preconditions to use this type of supervised learning algorithm is a large and classified dataset. Every new entry in this dataset will increase precision and accuracy of the computation model, since it is further used as a learning source by the working algorithm. Consequently, once there are enough input training sets, such model will improve in time and return better results for each subsequent entry. Decision trees, neural networks, regression and genetic algorithms are samples of ML supervised learning algorithms [16].

In [20] researchers list regression as a learning task within supervised learning. Through regression, a learning process can be performed, which enables learning by demonstration from a given data set. According to the general framework for supervised learning models, regression models also deal with a dataset consisting of past observations, for which both the value of the explanatory attributes and the value of the continuous numerical target variable are known [17]. In other words, both the input and output data are previously defined and even more the target (output) variable is always defined by some functional relationship between the input variables. If there is only one independent (input) variable, the regression model is referred as simple model, otherwise it is known as multiple regression model [17].

Linear regression algorithms are one of the most used statistical techniques for various predictive purposes, including ML supervised learning. These algorithms provide accurate estimation values for used coefficients, while minimizing the error produced as a difference between the real output value and the value received after each iteration [14], [18]. They can be used in a single input scenarios, as well as in scenarios where multiple input variables define the output. When the starting values for each coefficient are randomly selected and squared errors are calculated for each pair of input and output values, the linear regression model is known as gradient descent operation [18]. In such analysis, the learning rate for improvement step of determination must be defined in advance, to achieve optimal results [18], [28].

This study utilizes ML supervised learning with linear regression algorithm, more precisely gradient descent operation, as ML technique that is best used when the parameters cannot be calculated analytically [28]. Based on this algorithm, it presents a typical case study of business process implementation using ML advantages. This approach is further analyzed from different aspects: predictive possibilities, precision and accuracy benefits, corporative advantages, possible performance issues, etc.

3. RESEARCH METHOD

3.1 DESIGN AND MEARSURMENTS

Following the proposed methodology, we have implemented ML technique in a real-case business problem, while conducting an analysis of expected results and significance to its practical application. A starting key point is a decision whether the business problem can be approach via ML implementation. Many consider ML as a popular branch in the science and believe that every single problem should be resolved by implementing one of its algorithms. Hence, our research led us to conclusion that some prerequisites must be considered to reach a right decision. Therefore, we believe that the following preconditions must be fulfilled before ML approach can be applied for different business solutions:

- Basic data relationship between input and output cannot be clearly understood
- Business case cannot be described with behavioral statements and rules, and thus it cannot be translated into a program code
- Business case declaration can be clearly defined and specified with input and output parameters
- Business case supports large and comprehensive dataset
- The functional relationship between input and output data is constantly evolving and the solution requests real-time recalculation and fast adaptation
- Need for parallel architecture in a request for rapidly delivered and compact solutions.

We have researched a case study in large enterprise company in Macedonia that satisfies the stated preconditions, while analyzing a purchasing module of an application for ERP software, deployed with SAP solution. In our methodical approach, we have developed a learning algorithm, which tries to predict real-value outcome for a given array of inputs. More specifically, the chosen business proccess involves creation of a purchase order (PO) for a particular vendor or application users, while based on the PO input data the algorithm chooses a release strategy (output) for that particular PO. Currently, the creator of PO makes this decision, while the proposed solution will enhance the process and guide the creator, with a logical tool (learning algorithm) responsible for the output. Table 1 contains a list of input variables for such PO, with additional attribute (level of importance) for the ML implementation.

The examination of input parameters that are relevant for the PO creation was the first step in our ML implementation. Hence, we needed to select inputs that will be included in the calculations, which further can influence effectiveness of the solution. Misused inputs can lead to incorrect results and possibly endanger the whole project. During our research activities, we have reached the following conclusions: some of the parameters (e.g. Plant) have a constant value in all POs and thus we can consider them as obsolete for the solution. In addition, other parameters that provide information that has no applied meaning will discourse the calculations if ever used. This group is usually consisted of IDs with incrementing number values, most of them different for every PO (e.g. PO Number, Requirement Number etc.). Consequently, we compiled the first version of input parameters for the solution, with ten parameters in total. Each parameter is paired with appropriate numerical record. More precisely, the values for the chosen input parameters are derived from a special catalogue, in a numerical or alphabet notation, so the output can be calculate via the linear regression algorithm.

TABLE 1.
Input variables for PO creation

Input Variable Name	ML Relevance
Purchase Order Number	No
Order Type	Yes
Vendor ID	Yes
Document Date	No
Payment Terms	Yes
Item Number	No
Item Category	Yes
Account Assignment Category	Yes
Material Number	No
Short Text	No
PO Quantity	No
PO Order Unit	No
Delivery Date	No
Net Price	No
Currency	Yes
Material Group	Yes
Plant	No
Requirement Number	No
Requisitioner	Yes
Requirement Item	No
G/L Account	Partially, using the GL Acc. Group
Cost Center	Yes
Receiver details	No

3.2 PROPOSED ALGORYTHM AND ML SOLUTION

Since our ML implementation in a real-case business solution uses linear regression algorithm (gradient descent subtype), the aim of proposed computations is determination of coefficients vector that can create proper output form the input array. Having in mind that we have identified an input array with ten variables, ten coefficients should be derived that would be further used for calculation of the output variable. As a result, we have constructed the following ML algorithm for the chosen business process:

$$Y(x) = Q_1x_1 + Q_2x_2 + Q_3x_3 + Q_4x_4 + Q_5x_5 + Q_6x_6 + Q_7x_7 + Q_8x_8 + Q_9x_9 + Q_{10}x_{10} \quad (1)$$

The determination of the coefficients array (Q_1, Q_2, … Q_{10}) is a time consuming process that involves extensive calculations, use of matrices, their normalization using average and deviation values, transposition of matrices and creation of hypothesis coefficient values out of the processing. Therefore, we have submitted the proposed algorithm to such computational process in order to find the most accurate values for the coefficients array. Hence, this resulting array of coefficients should satisfy the Equation (1) for as much as possible data combinations (input and output pairs), provided by the system database.

Before starting with the calculation process, initial values must be assigned to the coefficients. Overtime, during ML execution, the accuracy of these values is expected to be improved. As long as the application is active, every single iteration should update the values of the coefficients. In other words, as long as the application is actively used by end-users, this solution will help the system accumulate not only historical data of passed activities, but also information of how this data is determined, while providing knowledge that can be applied in future determinations. Hence, the solution will help the employees perform their jobs more efficiently, which ultimately will provide increased business value based on business data assets used in the process.

The proposed ML business solution is divided in two distinct areas: (1) definition of the business tool (implemented ML algorithm) and (2) compiled version of the solution tool. Figure 1 depicts its components and data workflow. The application database is defined as a collection of actual and previously imported data inputs related with PO creation, while central part is the implemented ML tool that delivers most accurate relationship between the input data and outcomes. The result of these calculation generate coefficients list that is later used as starting point for every runtime data (actual input) while computing its outcome.

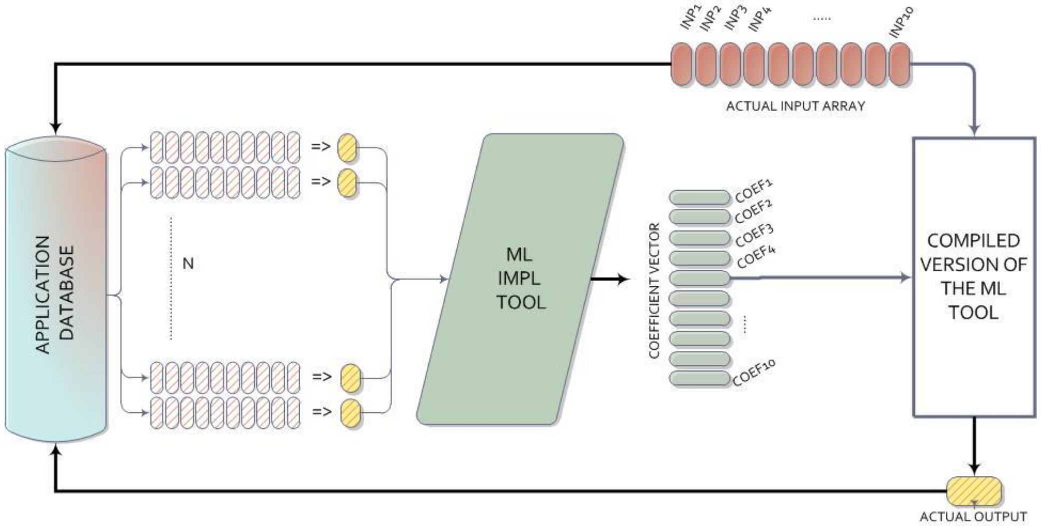

FIGURE 1. Schema for graphical representation of the implemented ML solution

One of the biggest challenges during development of the proposed ML business solution was the implementation part in the ERP system via programming language that could hardly be described as a full object-oriented language. In parallel, we have encountered several obstacles during execution phases because of highly demanding computational and processing consumptions. Therefore, we have excluded a real-time calculation of these coefficients from the final product and replaced it with one or two background executions per day that would not compromise processing performance during application usage. Later, we have used the received coefficient values for all PO creations and determination of release strategy for the given period.

4. RESULTS

After using the proposed ML solution in everyday activities for the chosen process within the company for 8 weeks, we were able to analyze its effectiveness in practice. Using the business data sets the ML solution was predicting an output release strategy, which was recorded and further compared with actual employees' decision, responsible for PO creation. During operation, the current employee was able to use the calculated ML release strategy, or decline the suggestion and use another release strategy, based on his experience. Therefore, we have gathered statistical data from the proposed ML business solution with the following information: (1) PO records created for a given period of time in correlation with the actual input parameters, (2) coefficient values for the same period of time, (3) proposed release strategies as ML outputs for the observed PO input data and employees' decision (actual release strategy), as well as (4) comparative statistical information.

Figure 2 shows comparison between the ML proposed release strategy (ML_Rel_Str) and actual release strategy by the employees (Act_Rel_Str), represented with values from the used catalogue, in a period of 8 weeks and 1135 POs. Similarly, Figure 3 illustrates the predictive effectiveness of the ML solution with deviations from the actual release strategy.

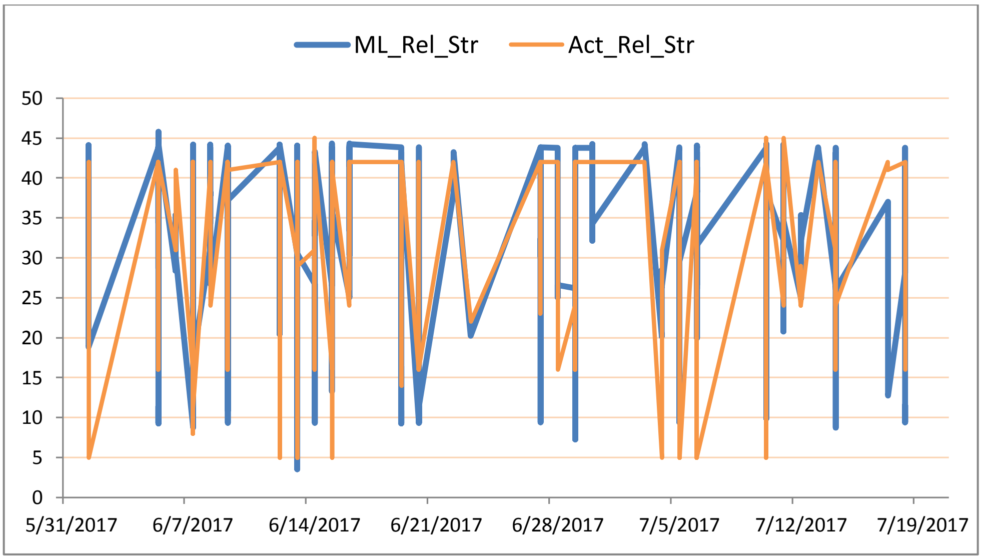

FIGURE 2. Comparison of ML predicted and actual release (output) strategy

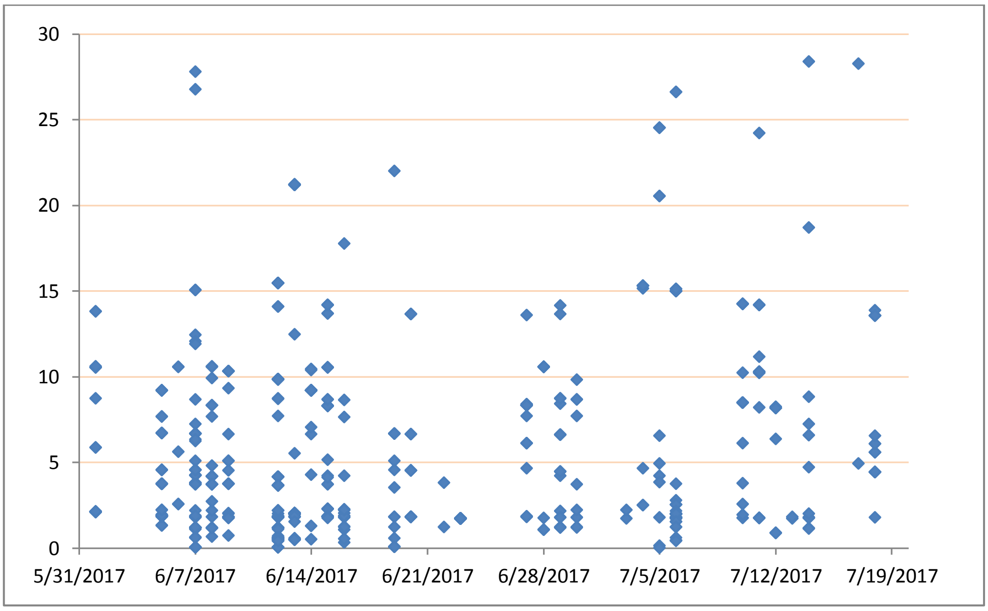

FIGURE 3. Deviation between ML proposed and actual output value

Even though the results were obtained in a short period of time, they confirmed the effectiveness of the proposed ML solution in a real business context. The comparison between ML predicted and chosen release strategy by the employees demonstrates similarity in both streams. The tendency of increasing or decreasing values in Figure 2 is nearly consistent in all resulting lines. Furthermore, the deviations in Figure 3 are mostly in the lower part, with decreasing density in the upper part of the axis. Hence, the number of ML predicted release strategies that closely match the chosen ones was significantly higher than inaccurate predictions (most of the predictions deviated from 0 to 15% of employees' decision).

In the first phase of its usage, this ML solution can help the researched organization streamline the PO release strategies based on the historical data, and thus provide competitive value to the business. Disregard of the specific employee, the organization was enabled to achieve consistent strategy within one of its core business processes and transform business information into knowledge. Ultimately, it can replace the human labor during decision making, which can be only use to supervise the business process and thus overcome subjectivity, wrong decisions and errors.

5. DISCUSION

The results from this study show that supervisory learning can be used as a business ML strategy. Although within the chosen process the exact inner relations of the large dataset were not clearly defined, the proposed ML solution was able to predict the chosen PO release strategy by the employees based on the desired output. It confirms the importance of ML classification indicated in [16], [23], since properly selected inputs in the ML calculations facilitated effectiveness of the solution.

It also supports the findings in [25], [26] that regression analysis will continue to be a valuable tool in business research. As in [18], this study verifies that even standard and most commonly used linear regression algorithms based on gradient descent operation can be beneficial. Such algorithm, when properly utilized in a ML solution, can transform business data assets into a benefit to the business process in the company. Having in mind the diversity of ML architectures and algorithms [27], the results have shown that the proposed solution provided optimal balance between computational complexity, amount of business data, performance and predictive output.

On the other hand, in [14] one of the major findings of this literature study is that predictive analytics are rare, since most of the related studies are more explanatory than predictive. Hence, this researches claim that statistically significant effect does not guarantee high predictive power, because the precision of the causal effect might not be sufficient for predictive accuracy, which can be further meaningful in practice. Therefore, one of the major contributions of this study is the predictive accuracy of the solution demonstrated in real-life implementation. The chosen algorithm is a statistical tool, but when properly utilized in a ML solution, it provided valuable predictive analytics in a business context. Although at the beginning the solution aims to streamline the business process and provide resilience to employees' inaccurate decisions, in the future it may automate the whole process and minimize the human error, consistent with future prospects of human labor as in [29], [30].

In the future work we will try to improve the prediction process, while refactoring the algorithm methods for execution-time minimization and allocation of memory, comparison of results between improved and older versions, etc. The chosen ML approach opens many possibilities within the researched company, which can be applied to other business processes, as well as can inspire other organizations to undertake similar initiatives in real contexts.

6. CONCLUSION

In this paper, we have presented a ML solution for enterprise business application, while discussing the significance of its implementation and derived benefits. We have identified prerequisites for proper ML approach and provided ML algorithm for a chosen business process, with encouraging results that can be meaningful to different companies while trying to improve their business processes. Hence, this study opens prospects for future research, as well as helps business decision makers understand the significance of machine learning for enterprise computing.

REFERENCES

[1] L. Bottou, "From machine learning to machine reasoning," *Machine learning*, vol. 94, (2) pp.133-149, 2014.

[2] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D.B. Tsai, M. Amde, S. Owen, D. Xin, "Mllib: Machine learning in apache spark," *Journal of Machine Learning Research*, vol 17, (34) pp.1-7, 2016.

[3] K. Langley, *Elements of machine learning*, Morgan Kaufmann, 1996.

[4] P. Jarvis, S. Parker, *Human learning: An holistic approach*, Routledge, 2006.

[5] M.V. Rosing, H.V. Scheel, A.W. Scheer, *The Complete Business Process Handbook: Body of Knowledge from Process Modeling to BPM, Volume I*, Elsevier, 2014.

[6] A. Ali, D.N. Jawawi, M.E. Yahia, "Using Naïve Bayes and bayesian network for prediction of potential problematic cases in tuberculosis," *International Journal of Informatics and Communication Technology*, vol.1, (2) pp.63-71, 2012.

[7] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of machine learning*, MIT press, 2012.

[8] A. Jain, H.S. Koppula, B. Raghavan, S. Soh, A. Saxena, “Car that knows before you do: Anticipating maneuvers via learning temporal driving models,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3182-3190

[9] S. Heinis, S. Kumar, S. Gezari, W.S. Burgett, K.C. Chambers, P.W. Draper, H. Flewelling, N. Kaiser, E.A. Magnier, N. Metcalfe, C. Waters, “Of Genes and Machines: Application of a Combination of Machine Learning Tools to Astronomy Data Sets,” *The Astro-physical Journal*, vol. 82, (2) p.86, 2016.

[10] H. Chen, R.H. Chiang, V.C. Storey, “Business intelligence and analytics: From big data to big impact,” *MIS quarterly*, vol. 36, (4) pp.1165-1188, 2012.

[11] K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012

[12] E.R. Sparks, A. Talwalkar, D. Haas, M.J. Franklin, M.I. Jordan, T. Kraska, “Automating model search for large scale machine learning,” in *Proceedings of the Sixth ACM Symposium on Cloud Computing*, 2015, pp. 368-380

[13] S.M. Pourhashemi, “E-mail spam filtering by a new hybrid feature selection method using IG and CNB wrapper,” *Computer Engineering and Applications Journal*, vol.2, (3), 2013.

[14] G. Shmueli, O.R. Koppius, “Predictive analytics in information systems research,” *Mis Quarterly*, pp.553-572, 2011.

[15] R. García-Martínez, P. Britos, D. Rodríguez, “Information mining processes based on intelligent systems,” in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2013, pp. 402-410, Springer Berlin Heidelberg.

[16] L. Maruster, *A machine learning approach to understand business processes*, Technische Universiteit Eindhoven, 2003.

[17] C. Vercellis, *Business intelligence: data mining and optimization for decision making*, John Wiley & Sons, 2011.

[18] O. Shamir, T. Zhang, “Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes,” in *International Conference on Machine Learning*, 2013, pp. 71-79

[19] M.A. Waller, S.E. Fawcett, “Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management,” *Journal of Business Logistics*, vol. 34, (2) pp.77-84, 2013.

[20] D. Garcia-Sillas, E. Gorrostieta-Hurtado, E. Soto-Vargas, G. Diaz-Delgado, C. Rodriguez-Rivero, “Learning from demonstration with Gaussian processes,” in *Mechatronics, Adaptive and Intelligent Systems (MAIS), IEEE Conference on*, 2016, pp. 1-6

[21] Z. Lodhia, A. Rasool, G. Hajela, “A survey on machine learning and outlier detection techniques,” *IJCSNS*, vol. 17, (5) pp. 271, 2017.

[22] P. Harmon, *Business Process Change: A Business Process Management Guide for Managers and Process Professionals*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2014.

[23] A.E. Maxwell, T.A. Warner, "Differentiating mine-reclaimed grasslands from spectrally similar land cover using terrain variables and object-based machine learning classification," *International Journal of Remote Sensing*, vol. 36, (17) pp. 4384-4410, 2015.

[24] H. Chen, H. Zhao, J. Shen, R. Zhou, Q. Zhou, "Supervised machine learning model for high dimensional gene data in colon cancer detection," in *Big Data (BigData Congress), 2015 IEEE International Congress on*, 2015, pp. 134-141

[25] L. Hopkins, K.E. Ferguson, K. E., "Looking forward: The role of multiple regression in family business research," *Journal of Family Business Strategy, 5*(1), 52-62, 2014.

[26] F. Harrell, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, Springer, 2015.

[27] M.I. Jordan, T.M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science, 349*(6245), 255-260, 2015.

[28] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, 2010, pp. 177-186, Physica-Verlag HD.

[29] H. David, "Why are there still so many jobs? The history and future of workplace automation," The Journal of Economic Perspectives, 29(3), 3-30, 2015.

[30] E. Brynjolfsson, A. McAfee, *The second machine age: Work, progress, and prosperity in a time of brilliant technologies,* WW Norton & Company. 2014.