Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 201 (2022) 679–683

Procedia Computer Science

www.elsevier.com/locate/procedia

The 12th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS)
March 22 - 25, 2022, Porto, Portugal

The Performance Benefit of Data Analytics Applications

Matthias Pohl[a,*], Daniel Gunnar Staegemann[a], Klaus Turowski[a]

[a] *Faculty of Computer Science, Otto von Guericke University, Universitatsplatz 2, 39106 Magdeburg, Germany*

Abstract

The idea of gaining new insights about business processes from data is driving companies to invest in the area of data analytics. In particular, the opportunities to improve typical business metrics such as revenue, profit, but also market share are pivotal for the motivation. In most data analytics projects, it is a challenge to determine to what extent the existing company-wide or generally accessible data pool is sufficient for such insights. Furthermore, the use of appropriate methods and algorithms for the analysis is an ongoing issue from the perspective of the data scientist, even if the question of suitable data has been resolved. Accordingly, this paper will outline how a benefit of data and algorithms on performance indicators through results of data analytics projects. A suitable setup for determining this influence will then be defined and demonstrated.

Keywords: data analytics, performance indicator, benefit analysis, sensitivity analysis

1. Introduction

Nowadays, for some companies, instead of physical goods or services, the business drivers are their own as well as customer-related data and its corresponding analysis. Often the public product or service is not designed for cost recovery (e.g. Facebook, Pinterest, WhatsApp) and the business profit is achieved from data-based recommender systems of user behavior (e.g. customer targeting). Data-driven business operations seem to have a positive impact on productivity and business performance [15, 28, 29, 13].

Value-adding results from business analytics projects are not guaranteed and can lead to an investment without return. Data analytics as well as so-called data science is necessary, because there is not a general best model that serves any use case [25]. Further, the identification of additional data sources is a challenge that requires economic justification. It also allows the assessment of offers from data service providers.

In order to be able to make statements about the usefulness of data analytics, a basic understanding of the influence of such project results on typical business indicators should be obtained. Another challenge in project planning is

[*] Corresponding author.
E-mail address: matthias.pohl@ovgu.de

1877-0509 © 2022 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2022.03.090

to assess the potential of a data analytics project in advance. Experiences from retrospective analyses of completed projects can facilitate this process [1] However, especially in small and medium-sized enterprises, these empirical values may not be available, so that a different approach is required for independent realization. As a result, it leads to the question:

How to determine the benefit of data analytics on business performance? (RQ)

In the following section, fundamental topics will be highlighted first. In addition to general approaches of data analysis, the cost-benefit analysis and statistical error analysis will be addressed. After a short description of the general methodological approach, the propagated concept of analysis will be presented. A final demonstration will show a first step of the evaluation. Further steps will be touched upon in the conclusion.

2. Foundations

2.1. Data Analytics

In consideration of the typical business term *data analytics*, a transition to data mining [6, 23] or so-called data science [17] is suitable. The typical approaches describe the data processing and analysis in order to obtain insights that are supposed to support the solution of a business problem. Predictive models are built and evaluated from the processed datasets used.

Another approach is the NIST Big Data Interoperability Framework [17], which additionally considers the scalability of data analytics in the case of extremely high data volumes (e.g., Big Data) and is referred to as *data science*. The overall process is divided into the sub-processes *data acquisition*, *data preparation*, *data analysis* and *action* as well as *visualization*, with the analysis of data playing the central role.

Data analytics can generally be divided into 3 categories [16, 11]. An overview of statistical properties and exploratory views are called *descriptive analytics*. The forecast of future values with the help of data mining methods and statistical modeling is summarised as *predictive analytics*. The further processing of the results obtained with optimization methods to address business problems is called *prescriptive analytics*.

The procedure models mentioned are mainly high level approaches or characterizations. A deeper consideration of the analysis of data leads to the concept of machine learning or statistical learning. Thereby, the distinction between supervised learning (e.g. classification, regression) and unsupervised learning (e.g. clustering, association rule mining) is established [7, 27]. These differ in the requirement that observations of the target variable exist (supervised) or that the target variable must be constructed or inferred (unsupervised). Further, mixed forms and variants can be selected project-specifically with taxonomies or ontologies that provide overviews of analytical models [9, 18]. With the reference to optimization problems, one can distinguish between minimization problems or maximization problems [12].

2.2. Cost-Benefit Analysis

A cost-benefit analysis is a systematic analysis of the benefits and costs of an issue or project [1, 14, 19]. The actual value is the difference between the costs and the benefits. In the paper at hand, the focus is on the consideration of the benefits of data analytics project by maximizing or minimizing business performance indicators. However, these benefits are associated with a risk, since their achievement is always uncertain. A risk analysis is also performed in the context of a cost-benefit analysis and is equated with an uncertainty analysis [14]. Another area to study effects on value are assigned to sensitivity analysis in business literature [2, 21, 3]. The objective aim is to investigate the effects of perturbations and fixations of input parameters and model structures on the output.

2.3. Statistical Error Analysis

The exploitation of the results of data analytics requires an evaluation of the obtained prediction models [23]. Typical metrics for the evaluation of categorical relations are Accuracy, Precision and Recall [26] or mixed measures, e.g. F1-Score [4]. All in all, these measures refer to the estimation error of prediction problems [5]. In metric number scales, deviation measures such as root mean squared error [5, 7] or mean absolute percentage error [10]. In scenarios of unsupervised learning [7], manual evaluation or the help of similarity measures is necessary.

3. Approach of an Analysis Concept

For construction of the analysis concept the methodological approach of Design Science is followed [8, 20]. In this approach, technological and organizational requirements are reflected by the fundamentals from the scientific knowledge base to justify a new theory or artifact for solving an identified problem.

The aim of the analysis framework is to measure and evaluate the benefit of different key elements of a data analytics project on the analysis target. The proposed approach should take the basic concepts into account (see section 2). In the following, the structure of the analytical framework is first described, followed by an explanation of the procedure.

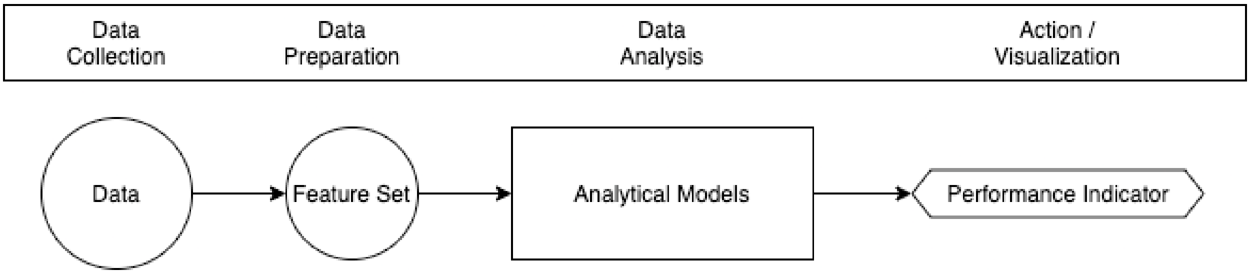

Fig. 1. Overview of the analysis concept.

3.1. Structure

Based on the phases of data analysis from the NIST framework, data collection, data preparation, data analysis and visualisation/action form the fundamental basis in the analysis framework. The analysis goal represents a *Performance Indicator* from the business context. Performance indicators can be interdependent so that conclusions can be drawn about other indicators. *Features* are generated from data that are used as input for data analysis. *Data* are provided by all data sources that can be made available in the data collection. Furthermore, new features can also be generated from meta-descriptions of data (e.g. simulated data).

Another influencing factor is the data analysis model (*Analytical Models*). A distinction can be made between descriptive, predictive and prescriptive models [16]. The influence of a model is represented by its output, which in turn can directly be a performance indicator. For example, the calculation of the moving average of a time series of sales data is a descriptive model as well as a performance indicator for the average sales revenue. The final business performance indicator represents a business-typical performance measurement.

3.2. Procedure

First, the framework parameters must be determined. In accordance with the procedure model from data analytics projects, suitable data must first be identified. In addition to the data, additional methods for data preparation must be selected in order to be able to generate data *features*. A pre-selection of *analytical models* must be made in the next step. It must be taken into account that data and models fit together.

The inputs are given by *features* and *analytical models*. During the analysis, in the first step, the different feature sets will be fixed and the results of several analytical models are calculated. Based on the compared results, the model with the best fit can be selected. The results of the model analyses can show which model has the greatest influence on the performance indicator. It should be noted that the model-specific evaluation metric (see section 2.3) may not be indicative of the impact on business performance.

In the second step, the feature set will be enhanced with simulated data features while the best analytical model is used for analysis. It is thus essential to identify the extent to which additional data have a further influence on the target variable. Minor differences could already be identified in the use of the different feature sets. The extent to which the values of the target variable can be improved can initially only be simulated. Thus, it can be discovered which datasets can be useful for the prediction problem.

4. Demonstration

In the following demonstrative example, a scenario of predictive maintenance from vehicle management is used. A logistics company has a fleet of trucks whose Air Pressure Systems (APS) provide sensor data for analysis. The APS supports the braking system and gear shifting system while the truck is in operation. The dataset[22] contains 171 dependent variables. Furthermore, information on the breakdown of the trucks is available so that a classification problem can be defined. The use case involves determining the failures of the trucks as accurately as possible. The quality of the prediction has an influence on the costs of maintenance.

The quality of the model is measured with a confusion matrix that leads to further evaluation metrics (see section 2.3). Correct predictions (true positives and true negatives) have no impact on the additional costs (*KPI*). However, false negatives lead to unnecessary maintenance, while false positives result in unexpected breakdowns. Unnecessary maintenance are valued at 10 cost entities, breakdown at 500 cost entities. The general objective is to minimise the cost. The benefit of the analysis can be seen directly in the performance metric *Cost*.

In the first step, three basic models are used for comparison. The models *Logistic Regression* and *Random Forest* are of simple form, while *Support Vector Machines* are decidedly more complex. The original dataset, a normalised or scaled set and a reduced dataset with the most informative data are used for the analysis. The results are shown in Table 1. The selection of the models used in the example leads to the result that *Random Forest* gives the best prediction in terms of *Cost*.

Table 1. Results of the first analysis step.

	Features	Analytical Model	Accuracy	Precision	Recall	F1-Score	Cost
I-A	original	Logistic Regression	0.97508	0.24855	0.20283	0.22338	85800
II-A	scaled	Logistic Regression	0.98808	0.75556	0.48113	0.58790	55330
III-A	scaled & sorted	Logistic Regression	0.98500	0.66000	0.31132	0.42307	73340
I-B	original	Random Forest	**0.99150**	0.89286	**0.58962**	**0.71022**	**43650**
II-B	scaled	Random Forest	0.99133	0.90299	0.57075	0.69942	45630
III-B	scaled & sorted	Random Forest	0.99000	0.81507	0.56132	0.66480	46770
I-C	original	Support Vector Machine	0.98233	1.00000	0.02147	0.04204	106000
II-C	scaled	Support Vector Machine	0.98275	**1.00000**	0.02358	0.04608	103500
III-C	scaled & sorted	Support Vector Machine	0.98333	0.87500	0.06604	0.12281	99020

5. Conclusion

The paper at hand presents a first approach to analyse the benefit of data analytics on business performance indicators. Taking into account the concepts of sensitivity analysis, it is shown that the choice of analytical models has a profound impact on the performance indicators. Although, the models lead to similarly good values of the evaluation metrics, the impact on the indicators (e.g. costs) is tremendous. In addition, it could be shown that there is potential for further improvement (e.g. reduction of costs).

The approach presented is designed broadly, thus, it needs to be further detailed in order to represent more specific model analyses. The demonstrating example presents rudimentary analysis problems. Other application scenarios from the logistics or production industry pursue far more complex issues (e.g. optimisation problems), which require detailed impact analyses. This will be taken into account in further development, as will the integration of other theoretical concepts, such as utility theory. Furthermore, value ranges for non-existent or latent variables are generated and integrated into the analysis. The aim is to determine to what extent the performance indicator can be improved. For this purpose, a probability distribution was derived according to the existing variables and new numerical series were created. The results of a new analysis with *Random Forest* and the extended data set range from 42630 to 48140 cost units. It shows potential for further systematic integration of this approach.

In the future, the Design Science approach should be pursued further and additional concepts should be integrated. Thereupon, a complete evaluation is also feasible [24].

References

[1] Boardman, A.E., Greenberg, D.H., Vining, A.R., Weimer, D.L., 2018. Cost-Benefit Analysis: Concepts and Practice. 5 ed., Cambridge University Press.

[2] Borgonovo, E., 2017. Sensitivity Analysis. volume 251 of *International Series in Operations Research & Management Science*. Springer International Publishing, Cham. URL: https://doi.org/10.1007/978-3-319-52259-3.

[3] Cacuci, D.G., Ionescu-Bujor, M., Navon, I.M., 2003. Sensitivity and uncertainty analysis. Chapman & Hall/CRC Press, Boca Raton.

[4] Derczynski, L., 2016. Complementarity, F-score, and NLP Evaluation, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Portorož, Slovenia. p. 6.

[5] Efron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316–331.

[6] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. AI magazine 17, 82–88.

[7] Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. Springer Series in Statistics, Springer, New York.

[8] Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design Science in Information Systems Research. MIS Quarterly 28, 75–105.

[9] Hilario, M., Kalousis, A., Nguyen, P., Woznica, A., 2009. A Data Mining Ontology for Algorithm Selection and Meta-Mining. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases , 76–87.

[10] Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688. URL: https://doi.org/10/dqmdbj.

[11] Lustig, I., Dietrich, B., Johnson, C., Dziekan, C., 2010. The Analytics Journey. URL: http://analytics-magazine.org/the-analytics-journey/. (accessed 11.11.2021).

[12] Marciano, A., Ramello, G.B. (Eds.), 2019. Encyclopedia of Law and Economics. Springer New York, New York, NY. URL: https://doi.org/10.1007/978-1-4614-7753-2.

[13] McKinsey Global Institute, 2016. The Age of Analytics: Competing in a data-driven world. Technical Report. McKinsey&Company.

[14] Mishan, E.J., Quah, E., 2007. Cost-benefit analysis. 5th ed ed., Routledge, London ; New York.

[15] Müller, O., Fay, M., vom Brocke, J., 2018. The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. Journal of Management Information Systems 35, 488–509. URL: https://doi.org/10.1080/07421222.2018.1451955.

[16] Naous, D., Schwarz, J., Legner, C., 2017. Analytics As A Service: Cloud Computing and the Transformation of Business Analytics Business Models and Ecosystems, in: Proceedings of the 25th European Conference on Information Systems, AIS, Guimaraes, Portugal. pp. 487–501.

[17] NIST Big Data Public Working Group, 2015. NIST Big Data Interoperability Framework: Volume 1, Definitions. Technical Report NIST SP 1500-1. National Institute of Standards and Technology. URL: https://doi.org/10.6028/NIST.SP.1500-1.

[18] Panov, P., Deroski, S., Soldatova, L., 2008. OntoDM: An Ontology of Data Mining, in: 2008 IEEE International Conference on Data Mining Workshops, IEEE, Pisa, Italy. pp. 752–760. URL: https://doi.org/10/fwrzxk.

[19] Pearce, D.W., 1983. Cost-Benefit Analysis. Macmillan Education UK, London.

[20] Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 24, 45–77. URL: https://doi.org/10.2753/MIS0742-1222240302.

[21] Saltelli, A. (Ed.), 2004. Sensitivity analysis in practice: a guide to assessing scientific models. Wiley, Hoboken, NJ.

[22] Scania CV AB, . Air pressure system failures in Scania trucks. URL: https://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set. (accessed 11.30.2021).

[23] Shearer, C., 2000. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of data warehousing 5, 13–22.

[24] Sonnenberg, C., vom Brocke, J., 2012. Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research, in: Peffers, K., Rothenberger, M., Kuechler, B. (Eds.), Design science research in information systems, Springer, Berlin. pp. 381–397.

[25] Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Philippines.

[26] Van Rijsbergern, C., 1979. Information Retrieval. Butterworth-Heinemann.

[27] Vapnik, V.N., 1998. Statistical learning theory. Adaptive and learning systems for signal processing, communications, and control, Wiley, New York.

[28] Vitari, C., Raguseo, E., 2020. Big data analytics business value and firm performance: linking with environmental context. International Journal of Production Research 58, 5456–5476. URL: https://doi.org/10.1080/00207543.2019.1660822.

[29] Wamba, S., Gunasekaran, A., Akter, S., Ren, S.F., Dubey, R., Childe, S., 2017. Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research 70, 356–365. URL: https://doi.org/10.1016/j.jbusres.2016.08.009.