EARLY DETECTION OF LOGISTIC PROCESS DELAY WITH MACHINE LEARNING ALGORITHMS

[1]Piotr JANKE, [2]Tomasz OWCZAREK

Silesian University of Technology, Faculty of Organization and Management, Gliwice, Poland, EU,
[1]piotr.janke@polsl.pl, [2]tomasz.owczarek@polsl.pl

Abstract

In this article an application of datamining algorithms for early detection of the delays in the process is presented. We used data of the real-world process of customer order fulfillment for industrial robots replacement parts in a multi-branch environment extracted from SAP system event log. Four different algorithms were tested: C&RT, boosted tree, random forest and artificial neural network in order to check their ability to detect if the whole process lasts longer than a specified time.

Keywords: BPM, process mining, data mining, machine learning, logistics process

1. INTRODUCTION

At present, there is a growing interest in the possibilities of modern IT tools in analyzing, monitoring and improving business processes. New solutions are created based on a process approach to the organization using data generated by the organization's information systems as a source of useful knowledge. Exploration of business processes (process mining) is now more often (apart from data mining) the technique used to discover knowledge about inflows of information (including those accompanying material flows) in enterprises [1]. This information enables optimization of business processes by means of static analysis of actually realized flows or as a source for systems of predictive monitoring of business processes [2]. Operating on data related to actual system events (event logs) is free from, above all, errors of subjective evaluation of what processes are being carried out. Machine learning methods are very often combined with the technique of process mining as its complement [1,3]. Therefore, the use of data mining techniques in the era of widespread computerization seems to be justified.

The article presents process exploration research carried out with the use of data gathered from an integrated IT system. The aim of the work is to assess the suitability of the data mining environment for the analysis of a selected logistics process.

2. METHODOLOGY

Here we simply describe our approach. For particular events extracted from event logs exported from ERP system we try to predict if the whole process is going to be delayed. So the response variable is binary categorical variable with two values: delayed and not delayed. As predictor variables we take diferencess between event dates (expressed as number of days between events) extracted from system event logs. Some data are treated as “historical” and are used to train the classification models.

2.1. DESCRIPTION OF THE DATA

In the research we used data collected in the SAP IT system of industrial robot distribution companies from 01-01-2016 to 31-05-2017. The generated report covers the course of the client's order for new spare parts. The process supported by the system integrated with the common database of the course includes in particular

the branch of the enterprise in Poland and the headquarters in Germany. The data structure contains information about **9** system events for each of **532** cases in the form of a time stamp (timestamp) of its completion.

Table 1 Types of tasks and their description in the process of fulfillment order [own study]

Task	Description of the activity
Order creation	Date of placing the order by the customer
GM - order to transfer	Date of the internal order
Service unit	Date of issue Packing List.
Transport	Date of issue of the consignment note
DM editions Mat	Date of the internal external issue
Confirmation of the service	Date of selection in the system of completion of the contract
Internal settlement	Date of invoice issue between branches
VAT invoice	Date of issuing the final invoice for the customer
Ending the order	Date of closing the order in the system

The description of system events describes the activities carried out in the process. The data does not contain precise information about the order of operations or the relationships between them. In total, the data contains information about the end of the duration of individual activities divided into individual instances of the process (cases) by numbers of subsequent documents generated from the SD (Sales and Distribution) module of the SAP system. Data prepared in this way, arranged in the form of a spreadsheet, constitutes the entry base for process exploration.

As predictors we used 6 variables representing the time differences between event dates and the date of the initial event "Order creation". The last event "Ending the order" was excluded from the calculation due to the possibility of calculating the duration of the process in a deterministic way. The "Confirmation of the service" event has been removed from the predictors due to lack of variation.

Table 2 Basic statistics of predictors [own study]

	Variable name	Type	Role	Mean	Standard deviation	Skewness	Kurtosis	Observed minimum	Observed maximum
1	act_gm_zlec_przen	Continuous	Input	5.2841	16.8501	4.8463041	28.88297	0	147
2	act_jedn_obs	Continuous	Input	0.0795	0.53315	11.192883	149.5506	0	8
3	act_transport	Continuous	Input	0.017	0.18411	13.493653	203.7437	0	3
4	act_dm_wyd_mat	Continuous	Input	0.0284	0.24873	10.453035	117.749	0	3
5	act_roz_wew	Continuous	Input	2.2045	0.96856	7.924568	89.58153	1	15
6	act_faktura_vat	Continuous	Input	6.7869	5.07129	0.8201166	0.567946	0	26
7	y	Categorical	Target						

All collected data was divided into two sets: training and test. The first dataset (352 cases) was used to train the chosen models and optimize their parameters. The test data set was completely disabled during the training phase and was used to assess the quality of the final model.

2.2. TRAINING AND EVALUATING

As the first model we tested Classification and regression tree algorithm. **Classification and regression tree (C&RT**) are used to determine the affiliation of cases or objects to the qualitative classes of a dependent variable based on measurements of one or more explanatory variables (predictors). Analysis of classification trees is one of the basic techniques used in the data mining [4]. **Table 3** presents the results of the model in the training stage: from the total number of cases of 352 instances of the process in the training group the C&RT model made one false-positive prediction (detected lateness when it was not there) and four times it did not detect the actual delay.

Table 3 Summary Frequency Table (Prediction) C&RT [own study]

	y	Prediction (FALSE)	Prediction (TRUE)	Row (Totals)
Count	FALSE	318	1	319
Column Percent		98.76 %	3.33 %	
Row Percent		99.69 %	0.31 %	
Total Percent		90.34 %	0.28 %	90.63 %
Count	TRUE	4	29	33
Column Percent		1.24 %	96.67 %	
Row Percent		12.12 %	87.88 %	
Total Percent		1.14 %	8.24 %	9.38 %
Count	All Grps	322	30	352
Total Percent		91.48 %	8.52 %	

Second algorithm was **Random forest.** It is a model that consists of a collection of voting trees generated through random selection of input variables [5]. **Table 4** presents the results of the model in the training stage: from the total number of 352 instances of the process in the training group the C&RT model made no false-positive prediction but in five cases it did not recognize the actual delays.

Table 4 Summary Frequency Table (Prediction) Random Forest [own study]

	y	Prediction (FALSE)	Prediction (TRUE)	Row (Totals)
Count	FALSE	319	0	319
Column Percent		98.46 %	0.00 %	
Row Percent		100.00 %	0.00 %	
Total Percent		90.63 %	0.00 %	90.63 %
Count	TRUE	5	28	33
Column Percent		1.54 %	100.00 %	
Row Percent		15.15 %	84.85 %	
Total Percent		1.42 %	7.95 %	9.38 %
Count	All Grps	324	28	352
Total Percent		92.05 %	7.95 %	

The third tested algorithm was **boosted trees**, which was developed as an application of reinforcement methods to regression trees. The main idea is to create a sequence of simple trees, each of which is built to

predict the rest generated by the previous ones [4]. **Figure 1** shows the optimal number of trees obtained during training phase. After about 80 number of trees model starts to show an excessive fit to the data and its prediction accuracy decreases.

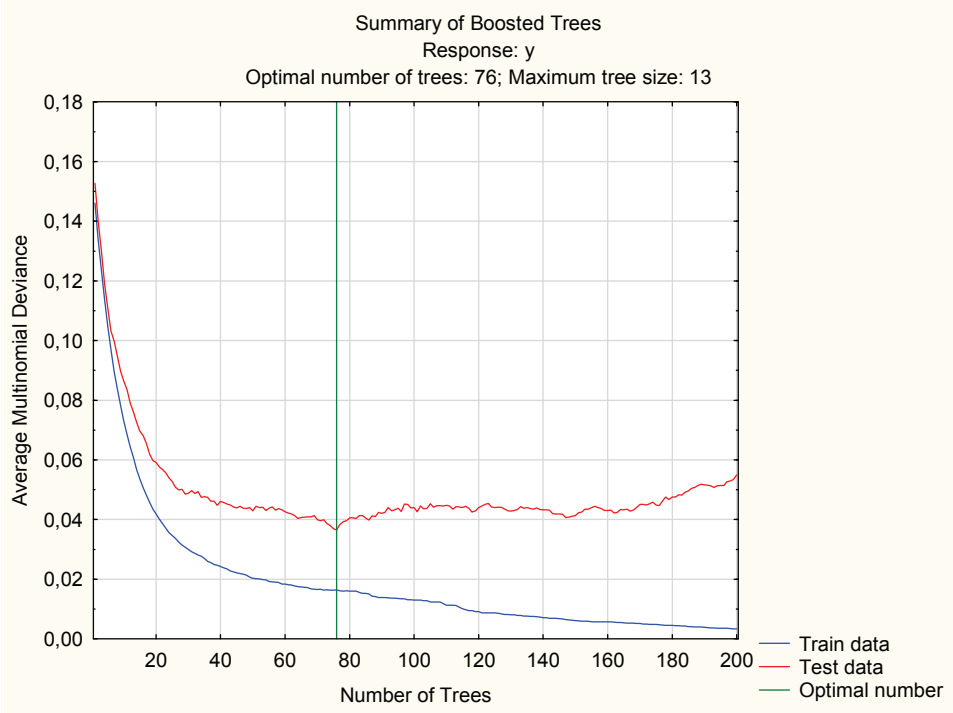

Figure 1 Average Multinominal deviance to number of Trees [own study]

Table 5 presents the results of the boosted trees model in the training stage: from the total number of 352 cases of the process in the training group the model made one false-positive prediction and missed one actual delay.

Table 5 Summary Frequency Table (Prediction) Boosted trees [own study]

	y	Prediction (FALSE)	Prediction (TRUE)	Row (Totals)
Count	FALSE	318	1	319
Column Percent		99.69 %	3.03 %	
Row Percent		99.69 %	0.31 %	
Total Percent		90.34 %	0.28 %	90.63 %
Count	TRUE	1	32	33
Column Percent		0.31 %	96.97 %	
Row Percent		3.03 %	96.97 %	
Total Percent		0.28 %	9.09 %	9.38 %
Count	All Grps	319	33	352
Total Percent		90.63 %	9.38 %	

The last machine learning method that we tested on our data is **artificial neural network** (we used a classic MLP networks with hidden layers from 3 to 10)**. Table 6** presents the results of the model in the training stage:

from the total number of 352 instances of the process in the training group the model made zero false-positive prediction but in nine cases it falsely predicted the delay.

Table 7 presents the summary of all models training errors. From the four tested algorithms the boosted trees was the best achieving 99,43 % accuracy (0,57 % error). The lift chart [6] of the models is shown in **Figure 2**.

Table 6 Summary Frequency Table (Prediction) Artificial neural network [own study]

	y	Prediction (FALSE)	Prediction (TRUE)	Row (Totals)
Count	FALSE	319	0	319
Column Percent		97.26 %	0.00 %	
Row Percent		100.00 %	0.00 %	
Total Percent		90.63 %	0.00 %	90.63 %
Count	TRUE	9	24	33
Column Percent		2.74 %	100.00 %	
Row Percent		27.27 %	72.73 %	
Total Percent		2.56 %	6.82 %	9.38 %
Count	All Grps	328	24	352
Total Percent		93.18 %	6.82 %	

Table 7 Error of the models [own study]

Model ID	Name	Training error (%)
3	Boosted trees	0.57
2	Random forest	1.42
1	C&RT	1.42
4	Neural network	2.27

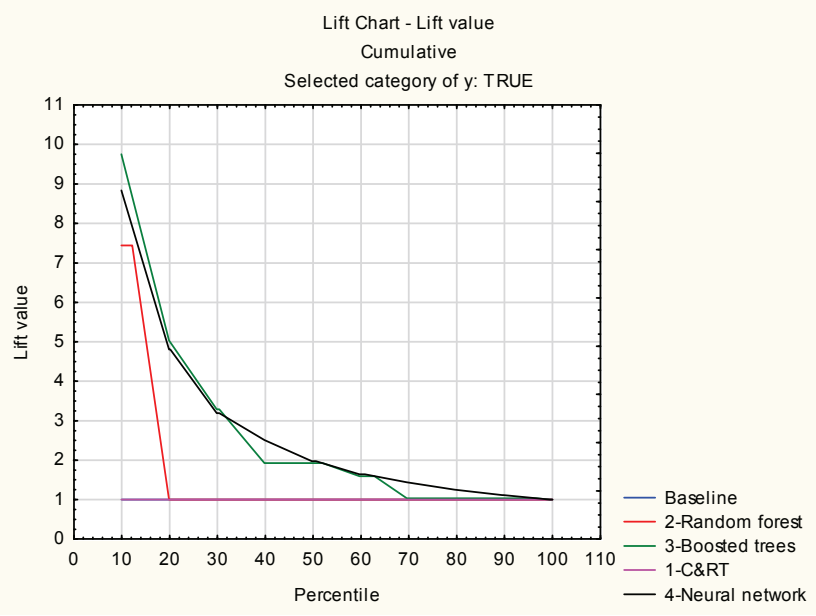

Figure 2 Lift chart graph comparing the tested algorithms [own study]

After completing the comparison of the algorithms, the boosted trees model was selected for final evaluation on the test dataset. From the total number of 171 cases it achieved an error rate of 1,16 % (two mistakes were made: one false-positive and one false-negative).

3. DISCUSSION AND RELATED WORK

From the four tested algorithms it was the boosted trees that scored the best, achieving 0,57 % error rate on train dataset, and 1,16 % on test dataset. It is worth to mention that these results were obtained without any deepened data exploration or preparation, with the exception of removing one variable with minimum variance. Proper exploratory analysis conducted before the training phase could further improve the results [7].

Similar approaches but with different algorithms can be found in the literature. Clustering methods were used by Folino et al. [8] to predict violations in service level agreement terms and by Kang et. al. [9] to detect abnormal termination. Metzger et al. compared three technics: neural networks, constraint satisfaction and Quality-of-Service aggregation in order to predict process outcome [10]. Their results suggest that various prediction techniques can combined in order to improve specific metric (e.g. to minimize false-negative prediction rate). Taking this into consideration it can be concluded, that the performance of different algorithms is strongly dependant on the modelled process and its data.

4. CONCLUSION

The aim of the article was to check the suitability of the data mining environment for the analysis of logistics process. We tested four data mining algorithms in order to assess their ability to detect process delay based on the historical data. The adopted approach assumed dividing the dataset into training and testing datasets - in order to "simulate" the historical and future cases. The results indicate that all four tested models have very low (below 3 %) training error. The best algorithm during the training phase turned out to be boosted trees and it was selected for the evaluation phase, where it obtained the result of 1.16 % error rate. It should be noted that the data used in the research were extracted from an ERP system and described a real-world process of clients ordering new spare parts from an industrial robot distribution companies.

The conducted research suggests that data mining classification algorithms are suitable for the analysis of logistics process and they can achieve very high accuracy in process delay detection. The results also show that the best algorithm is strictly dependent on the data - different models can be more appropriate in some cases, but less efficient in others. Choosing the best algorithm requires a proper approach.

REFERENCES

[1] VAN DER ALST, W.M.P. Process Mining - Discovery, Conformance and Enhancement of Business Processes. Berlin: Springer-Verlag, 2011. p. 352.

[2] MAGGI, F.M., DI FRANCESCOMARINO, C., DUMAS, M. and GHIDINI, C. Predictive monitoring of business processes. In 26th International Conference on Advanced Information Systems Engineering. Thessaloniki: Springer, 2014, pp. 457-472.

[3] R'BIGUI, H. and CHO, C. The state-of-the-art of business process mining challenges. International Journal of Business Process Integration and Management. 2017. vol. 8, no. 4, pp. 285-303.

[4] TIBCO Statistica Online Documentation [viewed 2018-10-23]. Available from: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Gxx/Gcrt/Overviews/ComputationalDetails

[5] BREIMAN, L. Random forests. Machine learning. 2001. vol. 45, no. 1, pp. 5-32.

[6] YEH, I.C. and LIEN, C.H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. 2009. vol 36, no. 2, pp. 2473-2480.

[7] OWCZAREK, T. An example of exploratory analysis for predictive business process monitoring, In 30th International Business Information Management Association Conference (IBIMA), Madrid, 2017, pp. 4224-4228.

[8] FOLINO, F., GUARASCIO, M. and PONTIERI, L. Discovering context-aware models for predicting business process performances. In OTM Confederated International Conferences "On the Move to Meaningful Internet Systems". Berlin: Springer, 2012, pp. 287-304.

[9] KANG, B., KIM, D. and KANG, S.H. Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction, Expert Systems with Applications. 2012. vol. 39, no. 5, pp. 6061-6068.

[10] METZGER, A., LEITNER, P., IVANOVIĆ, D., SCHMIEDERS, E., FRANKLIN, R., CARRO, M., DUSTDAR, S. and POHL, K. Comparing and combining predictive business process monitoring techniques. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2015. vol. 45, no. 2, pp. 276-290.