A C T A U N I V E R S I T A T I S N I C O L A I C O P E R N I C I
DOI:http://dx.doi.org/10.12775/AUNC_ZARZ.2020.1.005 ZARZĄDZANIE XLVII–NR 1 (2020)

Pierwsza wersja złożona 21.06.2020
Ostatnia wersja zaakceptowana 28.08.2020

ISSN (print) 1689-8966
ISSN (online) 2450-7040

Krzysztof Wojtkowiak[*]

DATA MINING ANALYTICS FUNDAMENTALS AND THEIR APPLICATION IN LOGISTICS

A b s t r a c t: The article describes several basic data mining fundamentals and their application in logistics and it consists of two sections. The first one is a description of different parts of data mining process: preparing the input data, completing the missing data, classification method using k-nearest neighbours algorithm with theoretical examples of usage conducted in open-source software called R and Weka. The second section of the article focuses on theoretical application of data mining methods in logistics, mainly in solving transportation problems and enhancing customer's satisfaction. This section was strongly influenced by data provided by DHL enterprise report on Big Data. The data used in theoretical examples is of own elaboration.

K e y w o r d s: logistics, data mining

J E L C o d e: L91

INTRODUCTION

Data mining may seem very complicated and complex subject at first glance. There are many tools to use, ranging from statistics to artificial intelligence, various types of databases, there are plenty of both free and paid software. However, it is not surprising as we look how rapidly the size of digital information is increasing nowadays and how many interesting patterns one could extract from it. Wal-Mart is a great example of this process, this company gathers data of over 20M transactions every day [Fayad, Piatetsky-Shapiro, Smyth, Uthurusamy; 1996, p. 38]. The extracted data is not only presented as numbers or words. For example, NASA's Earth Observing System is able to save dozens gigabytes of picture data every hour [Han, Fu, Wang, Chiang, Gong, Koperski, Li, Lu, Rajan, Stefanovic, Xia, Zaiane; 1996, p.2]. Due to algorithm usage, one is able to showcase findings that have not been discovered yet. Although,

[*] Contact information: Krzysztof Wojtkowiak, Nicolaus Copernicus University in Toruń, Faculty of Economic Sciences and Management, ul. Gagarina 13a, 87-100 Toruń, email: krzywojw@gmail.com

before one is able to present the results of data mining process, there are several conditions that must be met. Firstly, one has to make sure that input data is complete (there are no blanks), secondly one has to choose the best calculating method for each case. Last condition that has to be met is checking the visualization for mistakes and eliminating them. The main goal of this article is to showcase main, basic techniques of completing and extraction of data, tools that are used in these processes and their application in logistics using theoretical approach and created for this purpose scenarios.

1. DATA MINING FUNDAMENTALS

1.1. Preparing the input data and completing missing data

During data analysis process one can sometimes face several difficulties. Used data can be incomprehensible for a person that is conducting the research, there could be different units used or the database is simply incomplete. Figure 1 shows examples of errors that can be met during checking the input data.

Figure 1. Example of fictional bank's database

No.	Customer ID	Sex	Age	Transaction amount
5	12034	M	31	1052
6	C215	M	44	1111
7	11105	F	??	10654932
8	??	M	21	3
9	64393	N	UA	1300
10	F223H012	N	UA	??
11	62012	F	37	2000
12	??	F	54	17451

Source: Own elaboration.

After quick analysis of the data shown in this example, some questions considering the value and meaning of presented data arise. Customers ID with letters may originate from other source, so they may be foreign customers, for example. What does the letter "N" in "Sex" column mean? Does "UA" mean "underage"? Aside from these obvious questions, one can see that some brackets lack values and some have values that appear to be much different than the rest in this set, so there is a high probability that they are incorrect. In such cases, one has to choose from some of the most used solutions:

1. To replace unknown data with the fixed data that is consulted with an expert in particular field.

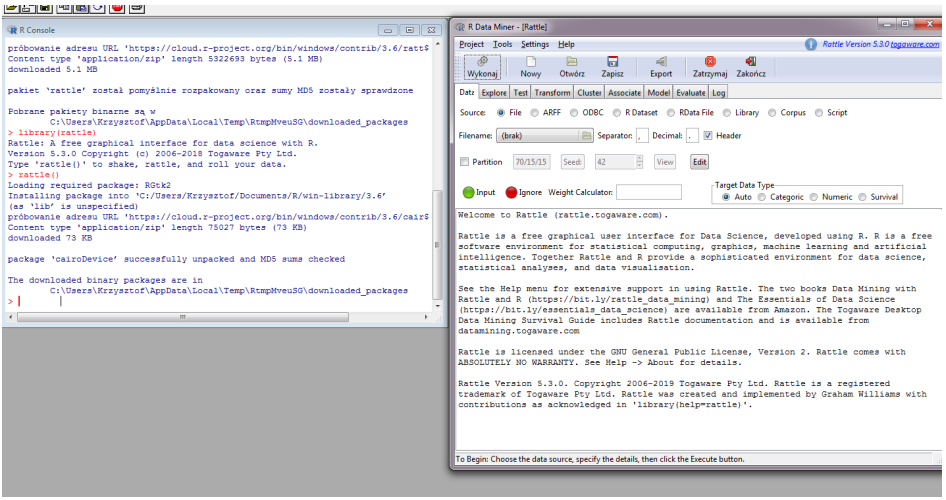
2. To replace unknown data with the mean value of the other data.
3. To replace unknown data with random values originated from probability distribution of the value.

The last solution seems to be the best for calculation purposes, however there is a substantial probability of obtaining unrealistic values, considering the matter of research so it forces researchers to think about the purpose of completing missing data.

Bayesian estimation theorem is a tool that might be helpful in such case [Larose, 2006, p. 8]. Thanks to this tool, one can answer to the following question: What value would be the most probable for the missing one, considering all other values?

In order to present those tools in theoretical approach, free programming language "R" was used. It has mainly statistical and visualization use. Except that, graphic user interface package called Rattle was used, as it enhances graphical presentation of results and further improves data visualization. The software was used to simplify calculations and optimize procedures. Figure 2 shows the starting screen of Rattle package when loaded in programming environment in R.

Figure 2. Starting screen of Rattle package when loaded in programming environment in R language



Source: Own elaboration.

After installing suggested plug-in which allows to read .xls files (Microsoft Excel), one was able to load example database which is shown in Figure 3. It includes data of 20 samples of example resource that were sent to a company and they were evaluated in following aspects: qualification test, durability test and cost-effectiveness analysis.

Figure 3. Scores of samples that were sent to the example company

No. of sample	Qualification test result (maximum = 100)	Durability test result (maximum= 5)	Cost-effectiveness analysis (maximum = 100)
1	66	1	13
2	80	3	
3	14	1	12
4	99	2	76
5		5	
6	51	4	88
7	32	4	
8		5	65
9	22	1	98
10	16	5	
11	80	2	34
12	78	3	55
13	66	5	44
14	34	1	54
15	28	4	
16	25		24
17		5	26
18	11	3	
19	5	5	28
20	76		79

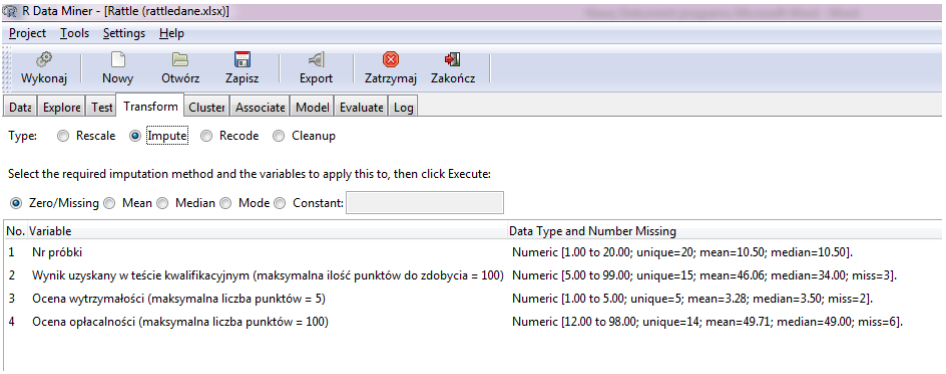
Source: Own elaboration.

It is easy to acknowledge that some of the values were missing, for example one does not know the result of qualification test of sample no. 5. Other missing values were highlighted in grey. After loading the data sheet in Rattle, program indicates number of missing elements in every class (in this example – in column) and describes them as "miss". The software offers several methods of completing missing data, all of which are shown in "transform" tab. Available options are:

1. Zero/missing – fills up empty spot with "0" mark.
2. Mean – fills up empty spot with the mean of data set in certain class.
3. Median – fills up empty spot with the median of data set in certain class.

4. Mode – fills up empty spot with the modal value of data set in certain class.
5. Constant – fills up empty spot with the fixed value that is chosen by the user. It can be either numeric value or letters.

Figure 4. „Transform" tab of Rattle software, number of missing elements



Source: Own elaboration.

For theoretical approach and as a example, the author has chosen the mean method for completing entries in qualification test and cost-effectiveness analysis columns and the median method for completing data in durability test column. Missing values were replaced, as displayed in Figure 5.

Figure 5. Completed data

00000	1.0	13.00000
00000	3.0	49.71429
00000	1.0	12.00000
00000	2.0	76.00000
05882	5.0	49.71429
00000	4.0	88.00000
00000	4.0	49.71429
05882	5.0	65.00000
00000	1.0	98.00000
00000	5.0	49.71429
00000	2.0	34.00000
00000	3.0	55.00000
00000	5.0	44.00000
00000	1.0	54.00000
00000	4.0	49.71429
00000	3.5	24.00000
05882	5.0	26.00000
00000	3.0	49.71429
00000	5.0	28.00000
00000	3.5	79.00000

Source: Own elaboration.

R programming language is only one many available ways of solving incomplete data problems, there are many more software on market to use. One have to be careful though, because it is very important to choose the correct completing method, in order to avoid unrealistic results.

1.2. Data exploration methods

Different visualization techniques are used during the work with the data, the choice usually depends on the future use of output data. Most popular methods are: regression, clustering, classification, discriminant analysis or association. This article focus mainly on classification method called KNN (K-nearest neighbours) however other methods are also briefly described.

Classification is about finding the best way to display data that is found in pre-built classes. The model is built based on provided data (an example of this kind of model is a decision tree or logical rules) which is later used to classify new objects added to the database or in order to better understand existing classes. For example, medical database may include rules that classify existing diseases in the database but also automatically classify every new added entries, which would diagnose patients too. Other known uses of classification is to find market trends or help with customer credit policy analysis.

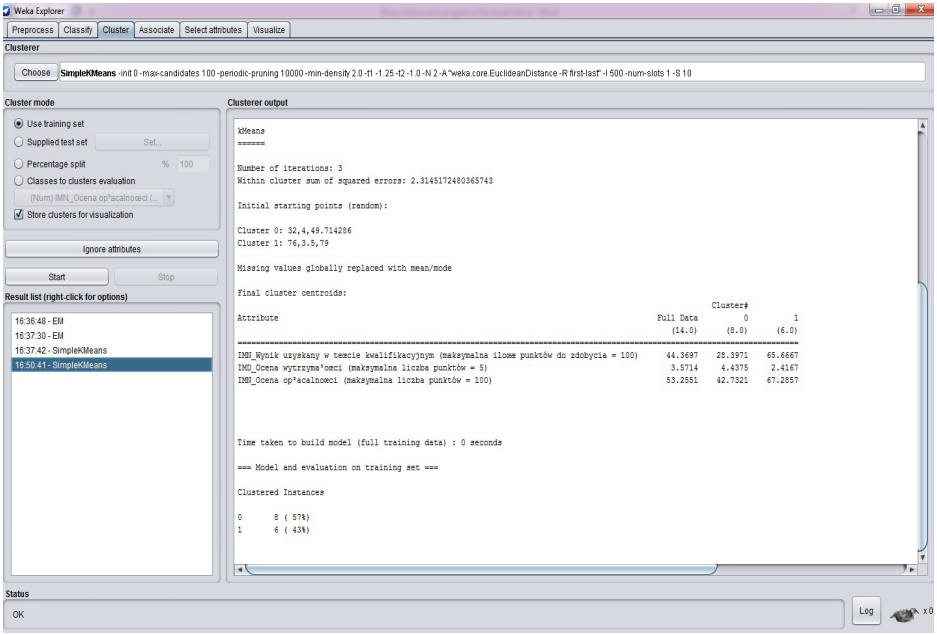
One can conduct KNN classification using free software called Weka which is used for data exploration. It was created by workers of Waikato University (New Zealand). In order to maintain order, the previous dataset from Rattle was used. Interesting options can be found in "explore" section, as shown in Figure 6.

Figure 6. Weka main screen

Source: Own elaboration.

In "cluster" tab one can choose the method of classification (in this example – KNN), select data source (in this case – database about resources sample used in Rattle) and proceed with the research. One can either choose some variables for analysis (percentage share) or conduct analysis based on other file, created only for this purpose. It is very important to select classes properly. In this case, one should ignore "No." class. The final effect is shown in Figure 7.

Figure 7. KNN classification method conducted in Weka



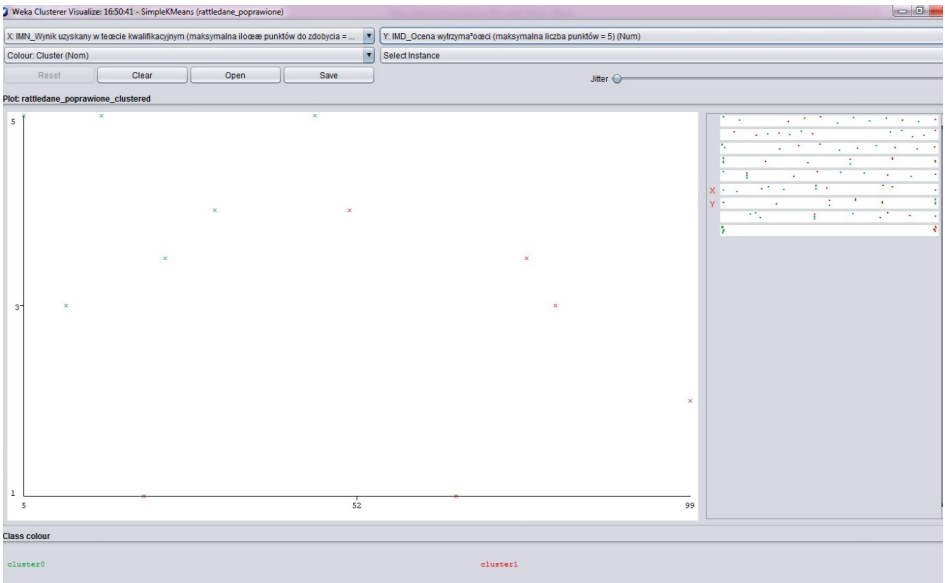
Source: Own elaboration.

The results that are shown in a text form are not always transparent and easy to work with so Weka software also enables visual representation of the results. Thanks to this particular option, one is able to notice that software classified input data into two groups (clusters) in red and green colour. Moreover, Weka's graphical visualization enables even more precise look into the classification of particular classes (columns) or allows to conduct another classification process. In Figure 8, one can see graphical representation of conducted KNN classification, where the X-axis represents sorted values of qualification test results of samples and Y-axis represents sorted values of durability tests results of given samples. The method classified all elements of first (green) group in the top left corner of coordinate system as shown in Figure 8.

Weka software is an example of a very complex software which includes many useful data analytics options and certainly it is the tool that can be recommended. Besides classifying, there are other ways of working with the data.

Regression analysis is a method of analysis for estimating the relationships between dependent variable and one or more independent variables provided from databases in this case. Example of regression analysis application is predicting the probability of patient's recovery or cancer detection based on data found in patient's medical database [Langer, van der Kwast, Evans, Trachtenberg, Wilson, Haider; 2009, p. 329-331].

Figure 8 – Graphical visualization of KNN classification method in Weka



Source: Own elaboration.

Clustering, also known as grouping, is the method of finding finite amount of sets that describe data. In other words, it is organization of collection of patterns into clusters based on similarities. It is very important to know the difference between clustering and discriminant analysis. In discriminant analysis, one is provided with pre-classified patterns and the goal is to label the new pattern using given patterns. Usually, the given labeled patterns are used to learn the descriptions of classes which are then used to label a new pattern. It is very often used for designing products for a narrow group of customers. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. Labels are associated with clusters also, but these category labels are data driven which means that they originate only from data [Jain, Murty, Flynn; 1999, p. 265]. Clustering can be used for solving theoretical issue of grouping unknown species of animals into known categories based on their features.

Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. It is used for finding the relationships between variables.

The main principle of associations is to find connections between occurrence of elements in given data sets. The most popular way of doing this is by so-called market basket analysis. It is analyzing customer buying habits by finding associations between the different items that customers place in their shopping baskets. The discovery of these associations can help retailers develop marketing

strategies by gaining insight into which items are frequently purchased together by customers.

1.3. The results of data exploring and methods of presenting them

The data that is extracted using exploration methods can be presented in many ways, both traditional and digital. In disciplines such as data mining or machine learning one can use neural networks, decision trees or association rules. The chosen method must be understandable, transparent which allows easy interpretation of the results. Solutions that fulfill this criteria are both decision trees and association rules.

Decision trees are graphical method of enhancing decision-making process. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending. An example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with this leaf. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated for the subtree rooted at the new node [Mitchell, 1997, p. 52–53]. In machine learning process, decision trees are used to extract knowledge from data set. During construction of the decision tree there are two key requirements that have to be met. The first one is that learning set should be varied so it would be possible to qualify all subsets. Secondly, variables of the output class have to be discrete, not continuous, in order to qualify them unequivocally. The example of the data used for building a decision tree is shown in Figure 9. The tree itself is shown in Figure 10.

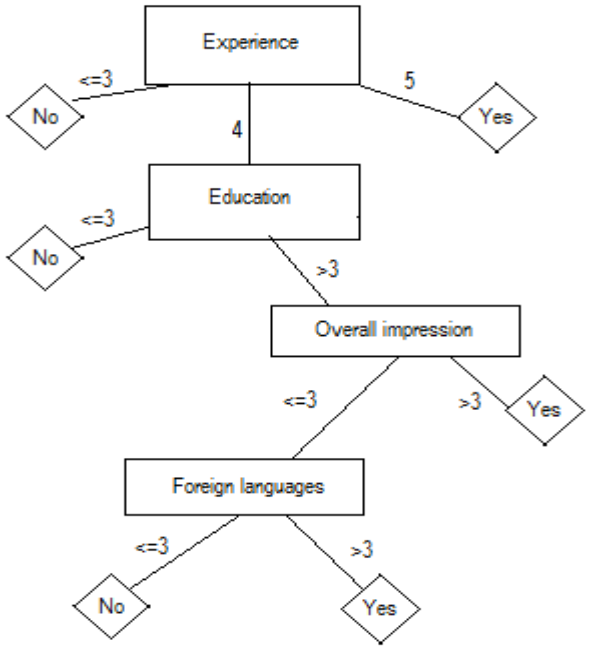
Figure 9 – Example of a data set that is used for building a decision tree

Age	Sex	Education	Foreign languages	Experience	Overall impression	Accepted
25	m	2	4	1	4	no
22	w	4	3	4	2	no
21	m	4	5	5	4	yes
29	m	1	3	2	3	no

Source: Own elaboration.

For theoretical purposes of this study the table with fictional scores of candidates applying for a job was created. The candidates were graded in 4 aspects (education, foreign languages, experience and overall impression), the grade 1 being the worse and grade 5 being the best score.

Figure 10 – Example of a decision tree based on previous table



Source: Own elaboration, based on http://edu.pjwstk.edu.pl/wyklady/adn/scb/wyklad12/w12.htm, [15.06.2020].

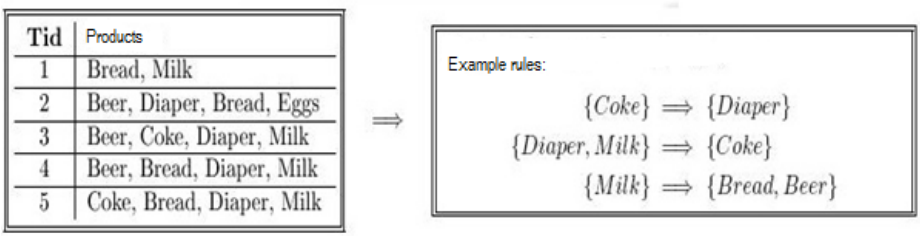
In this case, the experience was the most important factor of recruitment process so it had to be on top of the tree. From this point, every candidate that received 5 points in this category got the job, every candidate that received 4 point was directed to another stage of the process and if the experience score was 3 or lover the candidate was sent home. One can observe that it means that candidate no. 1 and 4 were sent home at this stage, candidate no. 2 was directed further and candidate no. 3 got the job at this stage. Because candidate no.2 scored 4 points in education category, she was directed to stage no. 3 where she failed to score above 3 point in overall impression category thus leaving the process. In conclusion, only candidate no. 3 was accepted for work. Despite meeting experience and education criteria, candidate no.2 failed in overall impression criteria.

Decision trees may become very big when there is a lot of data used and in this case they lose their transparency. In such circumstances using association rules is advised. They can look like this example:

R1 (a1, v1) ^ r2 (a2, v2) ^ … ^ RJ (aj, vj) → RK (ak, vk) ∧ RL (al, vl) ∧ ... ∧ RN (an, vn) where:
ai is the attribute,
vi is the simple value (e.g a number, a symbol) or complex value (set),
ri is the predicate (e.g an equation).

The association rule does not have predicted answer or expected value. Its purpose is to describe relationships between attributes. For example, from 1000 customers 200 bought milk and from 200 who bought milk 50 bought yogurt. Then then rule in this case would be: "If customer buys the milk then customer also buys the yogurt" with support 200/1000 = 20% and confidence 50/200 = 25%. Usually, there is the letter "S" used for support and "C" for confidence (Berry, Linoff; 2004, p.314). This kind of rules are applied in creating market baskets, that is how one can check how many customers who bought product A also bought product B.

Figure 11 – Example of created association rules based on products baskets.



Source: Own elaboration, based on: http://edu.pjwstk.edu.pl/wyklady/adn/scb/wyklad12/w12.htm, [15.06.2020].

2. DATA MINING APPLICATION IN LOGISTICS

2.1 Scientific approaches to data mining discipline in logistics

Logistics can be defined as a field of science that studies and develops solutions to create, optimize and control product, money and information flows to the final customer, providing the best possible service with reasonable costs. This processes include production location, warehousing, product transportation and deliveries [Beier, Rutkowski; 2003, p.13].

Data mining seems like an interesting discipline to connect with logistics. It can help with choosing the best facility location, use stock information to optimize warehousing of future deliveries, solve transportation problems, provide advice with choosing the best logistics operator for the company. Solutions like decision trees may be helpful with choosing the best way of transporting

products. Data mining may also be very useful in logistic customer service field. Information about bought products that can be found in feedback provided by customers after completing the order would be a great input data to use with data mining tools as it could easily predict customer's preferences which would be a great aid to marketing department. It could be even more serviceable in supply chain management. For example, discriminant analysis would support company management in choosing the best strategic partner. Clustering methods may allow better customer segmentation. All of these examples would ultimately lead to maximizing profits or decreasing the costs of a company. As useful as it may seem, the use of data mining methods in logistics is still mainly unexplored topic in scientific papers, probably because of the relative novelty of this discipline.

According to Fray da Silva and Praça [2015] that conducted a bibliometric study of 255 documents about data mining application in logistics there were 3 clusters identified that suggest different approaches to this topic. There were: theory development, market-oriented solutions and theory application.

As aforementioned study shows, theory development was the biggest cluster and it describes specific data mining techniques that can be used for specific problems like vehicle routing, site planning or container transportation. Cluster also includes techniques like genetic algorithms and heuristics. Example of a study in this group is a study conducted by Gen, Altiparmak and Lin (2006) where an extension version of two-stage transportation problem was considered to minimize the total logistic cost including the opening costs of distribution centers (DCs) and shipping cost from plants to DCs and from DCs to customers. To solve the problem, a priority-based Genetic Algorithm (pb-GA) was constructed, in which new decoding and encoding procedures were used to adapt to the characteristic of two-stage transportation problem, and proposed a new crossover operator called as Weight Mapping Crossover.

Another example of this category that was not mentioned by Fray da Silva and Praça [2015] can be a study conducted by Klepac [2014] about using data mining models for churn reduction and custom product development in telecommunication industries. Fuzzy expert system was used to create prospective customer value calculation model which helped to found most valuable telecom subscribers and better understand subscriber portfolio structure. Second tool used was time series analysis which provided predictors for predictive churn model which also helped to construct predictive churn model based on logistic regression. Third tool was a Social Network Analysis model which allowed to find the most valuable customers from the perspective of already existing subscriber network. Whole study helped Veza telecommunication company to lower the churn rate and allowed for a better understanding of a customer from the perspective of a Veza company.

Market-oriented solutions is a cluster where authors focused mainly on creating decision support systems and decision making models which were based on

the theories developed in papers from cluster no. 1. Most of the papers from this category are study cases and the main goal of them is to apply theory to practice. The first presented problem was about repair contract selection and provisioning problems in a factory which was solved by multicriteria decision-making model. Other confronted real life issue was improving iron ore quality during production stage, based on algorithms and simulation models. The author [Everett, 2001, p. 355–361] found possible solutions for improving quality in 2 companies and identified bottlenecks.

The third cluster was theory applications, cluster similar to aforementioned market-oriented solutions but without the dominant use of mathematical models or algorithms. It focused mainly on factors of companies' results or analysis of decision-making models, decision support systems. The given example was paper by Pan and Jang (2008) that described the usage of technology-organization-environment framework to discover the factors behind decision-making of adopting the enterprise resource planning (ERP) software in Taiwan's communications industry. The tool that was used was the regression analysis which resulted in showcasing model being able to explain 79% of the decisions made.

As aforementioned study shows, data mining does not only offer many tools that help to solve logistic problems but also there are at least 3 methods of approaching this discipline.

2.2. Cases of use in logistics - transport

According to study conducted by Paul, Saravanan and Ranjit Jeba Thangaiah [2011] third party logistics providers can save up to 10 – 40% on operational costs by making improved decisions about warehouse placements or transporting nodes. Therefore, there is a significant competition between these companies in logistics optimization. Data mining application strictly in transport is also worth mentioning as it naturally provides large amount of data (i.e. parcel tracking and inventory management) and can improve both on improving company result and worker's safety as classification on road accidents records can find that certain roads are more dangerous during the day because of fast-changing weather conditions in some areas. The study focuses on comparing 5 different approaches to the traditional linear programming and java-implemented programming models solutions of vehicle routing problem. Even though the results of traditional model and java model are identical, the second one is much faster therefore more optimized. Considering that the decision-maker still needs to choose from 5 different approaches, the time difference would be even bigger. To sum up, it shows the increase of company efficiency when using programming methods, often naturally connected to data mining.

A great example of the company that benefits a lot from implementing data mining is DHL. DHL is the world's leading logistics company. 380,000 people in over 220 countries and territories work every day to help customers cross borders, reach new markets and grow their business. The company is said to deliver 1,588,000,000 parcels per year (2019). Its document "Big Data in Logistics: a DHL perspective on how to move beyond the hype" includes many real life applications of data mining methods which allowed the company to develop further. In DHL, the benefits of using data mining are divided those connected to operational efficiency, customer experience and creating new business models.

The main idea behind increasing operational efficiency is the optimization of last-mile. This is a final stage in supply chain and very often the most expensive one. Optimization of this part lowers products costs therefore making it a promising application for data mining techniques. There are 2 different approaches to it. The first, revolutionary one is to process a massive stream of information to maximize the performance of conventional delivery fleet and is achieved by real time optimization of delivery routes. The second approach utilizes data processing to control new last-mile delivery model which allows to replace highly optimized workforce with just raw capacity of a huge crowd of randomly moving people.

Rapid processing of real-time information enables to solve travelling salesman problem, as it is remains the core challenge for last-mile delivery. Every vehicle in DHL company receives a continuous adaptation of the delivery sequence that includes geographical information, environmental factors and recipient status (i.e. from sensor-based detection or telematics databases). In this case, data mining provides correlated streams of real-time events for combinatorial optimization procedures which allow to re-route vehicles on the go by updating onboard navigation systems.

The second pillar of improving last-mile delivery is a crowd-based approach in a distribution network. The goal in to encourage people like commuters, taxi drivers or students to take on the last-mile delivery process on the routes that they are travelling and get paid for it. Increasing the number of such associates allow to take load off the delivery fleet. Despite the fact that crowd-based delivery has to be incentivized, it has potential to cut last-mile delivery costs, especially in cities. On the downside, a crowd-based approach also issues a vital challenge: The automated control of a huge number of randomly moving delivery resources. This requires extensive data processing capabilities, answered by big data techniques such as complex event processing and geocorrelation. A real-time data stream is traced in order to assign shipments to available carriers, based on their respective location and destination. Interfaced through a mobile application, crowd affiliates publish their current position and accept pre-selected delivery assignments.

Other aspect of improving last-mile delivery is to optimize utilization of resources. In DHL, good strategic and operational-level planning of distribution net-

works is the key to avoid supply shortages (which decrease customer satisfaction) and excessing capacities (which generates costs). Data mining techniques improve the reliability of planning and the level of detail achieved, enabling logistics providers to perfectly match demand and available resources.

In strategic network planning, big data (and so data mining) support the process by analyzing comprehensive historical capacity and utilization data of transit point and routes. What is more, these tools consider other factors like weather conditions or external economic information. In conclusion, thanks to big data and data mining, there is a much higher volume of information available for creating scenario modelling techniques and advanced regression.

The second large field of possible data mining applications in logistics is the customer service. The cost of obtaining a new customer is far higher than keeping the existing one therefore it is important to minimize customer attrition and understand customer demand.

The extensive use of data allows to identify valuable customers which may be on the verge of leaving the company using multiple data sources. By using techniques like clustering, semantic text analytics or natural-language processing it is possible to extract the attrition potential of every customer. Moreover, there are many triggers used, which initiate proactive counter-measures when loyalty indicators go dangerously low. Ultimately, high service level persists.

Working with a large volume of information may also improve CRM systems performance as big data can make a great use of data stored on public Internet sites. Aforementioned techniques like text mining or semantic analysis allow to extract precise data about even one product. This information can also be connected to region and publication time data which may allow to find a correlation.

The pillar of a good-functioning supply chain is a model describing all elements of supply chain topology and forces that affect the performance of the supply chain. To create this model, one need lots of aggregated and analyzed data. As the data stream is unstructured and continuously updated, big data analytics power the retrieval of input that aids detection of supply chain risks. Semantic analytics and complex event processing are used to classify patterns in such streams. When some pattern reaches critical point, the customer or supply-chain member is notified and is able to react. Equipped with this information, the customer can re-plan transport routes or ramp up supplies from other geographies.

Increasing popularity of big data therefore data mining methods benefits many small and medium-sized enterprises. With regression analysis, the fine-grained information in shipment database can significantly enhance the accuracy of traditional demand and supply forecast which is a great selling point for third parties. This extracted information is often used by SMEs that lack capacity to conduct their own research.

There is yet another way that SMEs can benefit from. DHL Address Management is a software that uses daily freight, express and parcel delivery data to verify addresses which may not seem like an issue in industrialized nations but it is a noticeable problem in developing countries. This verified, quality data may serve planning purposes for retail, banking and public sector entities.

CONCLUSION

Data mining is undoubtedly very interesting discipline which has a lot of potential. Significant development of IT technologies, Internet, data bases and overall data volume was the reason to start working on new tools that are able to process number of information that human is no longer able to do. It does not mean, however, that the human contribution is not needed anymore. Quite the contrary: the demand for qualified workers in this field is surging. Choosing the proper visualization method became even more important. There is a lot of free, open-source software that allow to introduce data mining everywhere one wants but also enables modifying tools to adjust to specialized needs.

Even though there are many pros of using data mining techniques, there are still some cons and requirements that have to be met. Firstly, in order to be able to use data mining IT department must cooperate with logistics which requires substantial change in a structure of organization, so before introducing these methods there has to be a mutual understanding of the challenges in the company. Secondly, data mining techniques require data to work with. In order to obtain such data, the company needs to maintain full transparency of information assets and ownership. strong governance on data quality must be maintained. Input data must be continuously cleansed from missing or corrupted values. Thirdly, data privacy must be granted. Even when a use case complies with prevailing laws, the large-scale collection and exploitation of data often stirs public debate and this can subsequently damage corporate reputation and brand value. Another important factor is human resources. Data mining requires qualified personnel, proficient in stats and mathematics as very specialized knowledge is needed to apply methods effectively. The companies interested in introducing big data should consider hiring more specialists in this field. Last, but not least – due to the rapid development of IT therefore some of the data mining techniques, one needs to acknowledge that, for example, method that was used successfully 5 years ago, may have better alternative now. For IT departments to implement big data projects there is a need for continuous evaluation of knowledge and developing new solutions.

BIBLIOGRAPHY

Beier F.I., Rutkowski K., (2003), *Logistyka*, Szkoła Główna Handlowa w Warszawie, Warszawa.

Berry M., Linoff G., (2004), *Data Mining Techniques For Marketing, Sales and Customer Relationship Management*, Wiley Publishing, Indianapolis, Indiana.

Everett J. E., (2001), *Iron ore production scheduling to improve product quality*, European Journal of Operational Research, 129/2.

Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., (1996), *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Massachusetts.

Han J., Fu Y., Wang W., Chiang J., Gong W., Koperski K., Li D., Lu Y., Rajan A., Stefanovic N., Xia B., Zaiane O.R., (1996), DBMiner: *A System for Mining Knowledge in Large Relational Databases*, Portland, Oregon.

Jain A.K., Murty M.N., Flynn P.J., (1999), *Data clustering: a review*, ACM Computing Surveys 31/3.

Jeske M., Gruner M, Weiß F., (2013), *Big data in Logistics, a DHL perspective on how to move beyond the hype*, DHL, Troisdorf.

Klepac G., (2014), *Data mining models as a tool for churn reduction and custom product development in telecommunication industries* [in:] Vasant P., Handbook of research on novel soft computing intelligent algorithms: theory and practical application, IGI Global, Hershey.

Langer L., Van der Kwast T., Evans A., Trachtenberg J., Wilson B., Haider M., (2009), *Prostate Cancer Detection With Multi-parametric MRI: Logistic Regression Analysis of Quantitative T2*, Diffusion-Weighted Imaging, and Dynamic Contrast-Enhanced MRI, Journal of Magnetic Resonance Imaging 30.

Larose, D.T, (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, PWN, Warszawa.

Mitchell T. M., (1997), *Machine learning*, McGraw-Hill Science/Engineering/Math.

Paul A., Saravanan V., Ranjit Jeba Thangaiah P., (2011), *Data Mining Analytics to Minimize Logistics Cost*, International Journal of Advances in Science and Technology 2/3.

Silva R.F., Cugnasca C.E, (2015), *What is the importance of data mining for logistics and supply chain management?* A bibliometric review from 2000 to 2014.

INTERNET SOURCES

https://www.r-project.org/ [access date: 15.06.2020]

https://rattle.togaware.com/rattle-install-mswindows.html [access date: 15.06.2020]

https://waikato.github.io/weka-wiki/downloading_weka/ [access date: 15.06.2020]

http://edu.pjwstk.edu.pl/wyklady/adn/scb/wyklad12/w12.htm [access date: 15.06.2020]

https://www.dhl.com/content/dam/downloads/g0/about_us/innovation/CSI_Studie_BIG_DATA.pdf [access date: 15.06.2020]