PERFORMANCE ASSESSMENT OF DIFFERENT CLASSIFICATION METHODS FOR COUPON MARKETING IN E-COMMERCE

Ľudmila PUSZTOVÁ, František BABIČ
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic,
E-mail: (ludmila.pusztova.2, frantisek.babic)@tuke.sk

ABSTRACT

E-Commerce environment represents typical commercial transactions that take place virtually online. UK Online Shopping and E-Commerce Statistics provided by Nasdaq estimates that by the year 2040, 95% of all purchases around the world will be performed through e-Commerce. In 2021, there will be 2.1 billion digital buyers worldwide, up from 1.66 billion in 2016 (Spiralytics). In 2019, the 31 billion digital coupons were redeemed up from 16 billion in 2014; 77% of consumers spend 10-50$ more than anticipated when redeeming mobile coupons (Invesp). These statistics and expected trends in the future motivated us to investigate a performance assessment of different classification methods for digital coupon marketing. For this purpose, we used available data provided by the Data Mining Cup 2015. We compared three methods for decision tree generation (C4.5, C5.0, Random Forest), Naive Bayes, Support Vector Machine, and Logistic Regression. We tested their performance within different data samples created from the initial dataset. The best accuracy was provided by the C5.0 algorithm (91,3%) and Support Vector Machine (91,5%). These machine learning methods also had the highest success rate for the other metrics.

Keywords: *digital coupons, e-commerce, classification, decision tree, Naive Bayes, Support Vector Machine*

1. INTRODUCTION

Nowadays, business success is not in the ability to make something, but in the art of selling it. This will show the difference between a successful and unsuccessful company. Sales promotion has an important place in a business, but many people mistake it with advertising. Both have a common goal, but their character is different. While the ad advertising says, "Buy our product," sales promotion calls for "Buy me now." Sales promotion involves offering preferential prices to the buyer, such as coupons and discounts. Coupons play a significant role in everyday life. Customers can get coupons from a variety of sources and use it in stores, such as newspapers e-mails, or websites. In 2011, the top 5 coupon distributors used channels like newspapers (89.4%), directly in the store (4.2%), direct mail (2.3%), magazines (1.5%), and product packaging (1.3%). The predominance of redemption coupons is maintained by the US, which in 2015 was redeemed by consumers of $ 127 million.

These statistics confirm the importance of this type of marketing within the e-Commerce environment and in the era of Big data also the usage of suitable analytical methods to support it. The coupons provide significant benefits not only for the customer but also for the seller. Sellers can secure customer returns and their repeated purchases; use them in marketing and promotional campaigns as an essential supporting tool. They can use the coupons to research the price sensitivity of different target groups (by sending them with different values). The rule of thumb says that customers who collect and use coupons are more price-sensitive than those who don´t use them. They can easily track coupons according to the repayment rate and the redemption point. All these approaches produce the data. And this data represents an important source for analytical purposes to answer the questions like who will make the purchase even without using a coupon? Who typically uses coupons? Or which customers will return? The answers can lead to a better understanding of customer segments and improving seller's decisions regarding maximizing profits and minimizing marketing costs.

The paper is organized as follows. Firstly, we introduce the relevant state of the art in the field of coupon redemption and used machine learning methods. Secondly, we briefly describe the CRISP-DM methodology and techniques used in the process. Thirdly, we described the whole analytical process on available data. In conclusion, we summarize the results and obtained experiences.

2. STATE OF THE ART

E-Commerce data represents an interesting source for various analytical purposes, such as a prediction of future buying behavior, customer profiling, market basket analysis or analysis of coupon marketing effectiveness. We briefly presented some case studies related to the use of various machine learning methods on transaction data.

Daskalova et al. analyzed the date from an extensive marketing study and an online survey realized with 1 thousand individual respondents as a representative sample of the American population [1]. The investigation focused on email coupon marketing. They found that 92% of American adults have received a promotional email and 65% have used one in 2 weeks. Next, the authors investigated their findings deeper and conducted an online survey with 151 participants with various backgrounds, such as age between 19 and 67, 47% female, lived in 36 different countries. Based on both results the authors identified four common coupon usage behaviour: sharing, saving, unsubscribing, and using.

The „O2O Coupon Usage Prediction in Daily Life" challenge hosted by Alibaba in TianChi platform, focuses on the factors that can affect customers' coupon usage behaviour. The dataset contains online, and offline user's purchase behaviour records from January 1, 2016, to June 30, 2016. Jianwei He and Wenjun Jiang [2] described their

approach on how to solve this task. At first, they conducted extensive data analysis and study users' coupon usage behaviours for predicting coupon usage probabilities. Secondly, they extracted some types of features that can significantly impact customers' behaviour because the useful feature can also help the merchant to make a good discount strategy, and the merchant can issue the coupon to the user who is more likely to use the coupon. They used a variety combination of features to train the model and observed its AUC score, e.g. Support Vector Machine, Random Forest, Gradient Boost Decision Tree, XGBoost, Naive Bayes, and observe the impact of each feature on the final prediction result. The authors converted the coupon usage probability prediction problem into a binary classification problem (the record in which the user receives the coupon and uses the coupon they marked as a positive sample; the record in which the user receives the coupon but does not use the coupon they marked as a negative sample). In some cases, they combined multiple models in different ways. For evaluation, the classifier was used precision, recall, and AUC score. Their result showed that XGBoost has the best performance, and Random Forest has the worst rating.

Wu et al. [3]focused on future coupon usage prediction and used data from this competition. After initial data analysis, they extracted features of users, merchants, coupons, and user-merchants from original data, and removed not relevant or redundant attributes (user number, merchant number, etc.). Experiments on data were built by machine learning methods such as Naive Bayes, k-Nearest Neighbour, Logistic regression, Neural network, Decision trees (CART), and Random forest (RF). Samples randomly divided into ten groups on average and done experiments using the methods mentioned above. Besides, data of each group, they divided into training and testing sets. For model evaluating, they used confusion matrix and inferring metrics like precision, recall rate, F1-measure, accuracy, and ROC curve. By comparing each metric for each algorithm, they evaluated that the RF and CART algorithms have better performance than the others. From the side of the ROC curve, RF was also the highest area.

We mentioned some of relevant and interesting studies in our previous work [4] focused on the same data. However, we have modified our goal and enriched experiments with new algorithms and evaluation metrics.

3. METHODS

We performed our analytical process in line with the CRISP-DM methodology typically used in the domain of data analytics [5]. This methodology defines six main phases, specifically business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

3.1. Machine learning methods

Decision trees are the most popular form of classifier representation, especially for their easy-to-understand representation of acquired knowledge [6]. Decision trees can handle high dimensional data and, in general, has good accuracy. In data mining, decision tree structures are a common way to organize classification schemes. For learning decision trees have been developed many algorithms [7], but we decided to use the most popular of them - C4.5, C5.0 and Random Forest.

The C4.5 algorithm used normalized information gain for splitting. This algorithm gets smaller decision trees and can give ruleset as an output for a complex decision tree. The great benefit is handling both numerical and categorical attributes [8][9]. C4.5 uses the concept of information gain for measuring purity. The information gain, Gain (S, A) of an attribute A, relative collection of examples S, is given by the equation: Gain (A) = I (S1, S2, · · ·, Sm) − E (A). In other words, gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

The C5.0 algorithm represents an improved version of the C4.5 that offers a faster generation of the model, less memory usage, smaller trees with similar information value, weighting, and support for the boosting. C5.0 algorithm uses the boosting algorithm for increasing accuracy and the concept of entropy for measuring purity. The entropy of a sample of data indicates how mixed the class values are; the minimum value of 00 indicates that the sample is entirely homogenous, while 11 shows the maximum amount of disorder. The definition of entropy can be specified as follows [10]:

$$E(S) = \sum_{i=1}^{c} -p_i log_2(p_i), \quad (1)$$

where E(S) means the purity of a specific data segment or the whole dataset, c represents the number of class labels, and pi reflects the proportion of observation within a class.

The Random Forest (RF) algorithm is a modification of bagging and builds a set of de-correlated trees. RF is a compound classifier used for classification and regression tasks, that average the results of multiple decision trees without pruning. The individual tree models must be independent, so random attribute selection is using for each tree. Since trees are separate, it is convenient and easy to process them in parallel [11]. The result of the classification is determining by voting. Random selection of the training set of each tree allows it to be validated on data that are not selected for its training ensures rapid validation. At the beginning, it is necessary to define some parameters like a few trees in the model and the number of randomly selected attributes in each tree. When selecting a test attribute in a tree node, *m* attributes are considering the total number of p attributes. At the next node (sub-node), only *m* attributes are considering again. But it is a different subset of m attributes than in the previous case, and yet, one attribute is selected. A strong predictor is using as the test attribute [12].

The Naive Bayes (NB) is one of the most straightforward classification techniques, but it is often one of the relatively accurate predictive methods [13]. Despite its simplicity, the performance is usually comparable to other more sophisticated approaches. Bayesian classifiers predict the probabilities that an example belongs to a class. They are based on the determination of the conditional probabilities of individual attribute values for different classes and the assumption that the attributes are independent of each other. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of

calculating posterior probability P(c|x) from P(c), P(x), and P(x|c), as we can see below [14]:

$$P(c \mid x) = \frac{P(x|c)P(c)}{P(x)}, \quad (2)$$

where P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes); P(c) is the prior probability of class; P(x|c) is the likelihood which is the probability of predictor given class, and P(x) is the prior probability of predictor.

Logistic regression is a prediction model for categorized quantities. Independent variables can be both numeric and categorized. Based on the generated model, it is possible to make predictions for unknown cases, including estimation of the probability of occurrence of individual target categories. Binary logistic regression [15] is a specific type of regression analysis in which the dependent variable is nominal and takes two values, usually coded as 0 or 1. Logistic regression is mathematically expressed as follows:

$$\log it\,(p) = ln\left(\frac{p}{1-p}\right), \quad (3)$$

where the term in parenthesis (probability ratio) is called odds ratio.

Support Vector Machine (SVM) algorithm is used for classifying tasks and typically provides the most accurate results compared to all other algorithms in terms of predictive accuracy. The absence of a local minimum is one of the main features of SVM. The SVM model is a representation of training data, and you can extract the densified data set with support vectors. The main task of the algorithm is to find a separating line (optimal separator) that maximizes the space between two different classes (which means correctly separate two different classes) and minimizes the upper limit of generalization error [16]. The separating line must be furthest from the training examples, while it is needed to calculate all distances of the training examples from the separating line. The smallest distance is called the border. The distance from the separating line to the border is called the range, so the classifier with the most significant range value will be more accurate.

The confusion matrix is one approach for the models and methods evaluation, that represents correctly and incorrectly classified records. Based on this matrix, we can determine whether a coupon was redeemed or not. This matrix is useful for measuring some metrics such as accuracy, classification error, sensitivity and specificity values, and AUC value.

The percentage accuracy of the classification, which represents the percentage of correctly ranked examples, is defined as:

$$Accuracy = TP + TN/(TP + FP + FN + TN), \quad (4)$$

as well as the classification error, defined as:

$$classification\ error = 100\% - Accuracy. \quad (5)$$

ROC curve captures the behaviour of the classification rate when varying the classification threshold via a graph. It is a commonly used method of visualizing performance in binary classification. The AUC value represents the area under the curve and quantifies the overall ability to distinguish between correctly and incorrectly classified cases. If the AUC is closer to 1, the classifier is more successful; if it is closer to 0.5, the classifier is very poorly successful. This curve plot has two parameters: sensitivity and specificity. Sensitivity (True Positive Rate) measures the proportion of actual positives that are correctly identified and is define as:

$$Sensitivity = TP/(TP + FN). \quad (6)$$

Specificity (False Positive Rate) measures the proportion of actual negatives that are correctly identified and is define as:

$$Specificity = TN/(TN + EP). \quad (7)$$

4. ANALYTICAL PROCESS

Our goal was to find out which selected machine learning method is the most suitable for transactional data analysis on customer purchasing behaviour. We tested their performance within different data samples created from the initial dataset.

4.1. Business Understanding

The business objective was to improve the effectiveness of related coupon marketing through an identification of different buying behaviour models. In the language of numbers, it means to better understand the customers and to increase the profit. The result could contain information like which brand of goods is the bestselling; which customers create orders without using a coupon; which customers buy premium products; which customers will use all three coupons, etc. It could help the seller to make decisions about proving coupons and maintaining the number of regular customers.

We transformed this business objective to the data mining perspective. It means that we applied the selected machine learning method to pre-processed data samples in order to determine which method will have the best performance in line with the metrics.

4.2. Data Understanding

The data came from the DATA MINING CUP 2015 competition [17], divided into two datasets. The both samples contained 6 722 orders from an anonymous online shop; the training set 6 053 and the testing 699. The orders are described by 32 variables, that focused on each coupon and the order and contains three target attributes (coupon (1,2,3) Used) representing the coupon redemption for three possible products in one order. The detailed description of this data sample can be found in previous work [4].

The distribution of values for target attributes is unbalanced. The value "0" means that coupon was not

redeemed and value "1", that coupon was redeemed. The prevailing value in all three cases is "0".

Training sample contained the orders from 2 961 different users, so one user could make more than one purchase. The highest number of purchases made by one user was 30, but most of users made only one purchase. Values of the attribute couponID indicate that some coupons are used multiple times.

The most common combination of all three target attributes was the combination like coupon1used-0, coupon2used-0, coupon3used-0). It means that most orders didn´t contain any coupon redemption. The least popular combination was the redemption of coupons 2 and coupon 3 without coupons 1. From the comparison of the couponID1 and brand1 attributes, we found that each coupon is determined for only one brand. This property also applied to coupon 2 and coupon 3.

An important part of this phase was the analysis of relations between attributes. For numerical attributes, we used correlation to each of the three coupons. The highest correlation was between attributes price1 and reward1 (-0,147); price2 and premiumProduct2 (-0,147); price3 and premiumProduct3 (-0,195). Other numeric attributes for all three coupons had a minimal correlation, i.e., the attributes were statistically independent.

4.3. Data preparation

In this phase, we focused mainly on new variable creation, metadata extraction, removing missing values, selecting attributes and data transformation. Our motivation was to provide more interesting information for the classification algorithms.

New variables:

- expensive/cheap_product_1,2,3 (created by establishing the boundary between cheap and expensive product);
- high/low/no_discount_product_1,2,3 with binary values (created by the difference between the price^i and basePrice^i attributes, then determining the limits for high, low and no discount);
- number_of_coupons_used (describes the number of coupon redemption for the relevant order);
- combination_of_used_coupons (we determined what different combinations of coupon usage can occur and then replaced these combinations with a nominal value (range from 1 to 8));
- diffTime (difference between time of couponReceived and orderTime divided into 8 groups in minutes).

Metadata extraction

- variable couponReceived divided into four new: receiveMonth, receiveDay, receiveMinute and receiveHour;
- variable orderTime divided into four new: orderMonth, orderDay, orderMinute and orderHour;
- priceDifference1 (difference between the price1 and basePrice1), priceDifference2 (difference between the price2 and basePrice2), priceDifference3 (difference between the price3 and basePrice3).

Data transformation

- variables brand1,2,3 transformed to nominal, i.e., we calculated multiplicity for each different brand and labelled them with new values. The new range was 0 to 26 for brand1, 0 to 26 for brand2, and 0 to 27 for brand3. If some concrete brand was the same in the attributes, it was labelled with the same new value. Similar approach we applied on variables categoryIDs1 (new range from 0 to 14), categoryIDs2 (0 to 16), categoryIDs3 (0 to 16), productGroup1 (0 to 174), productGroup2 (0 to 179) and productGroup3 (0 to 206);
- values in variable receiveHour replaced by three new: divided into three new: morning, afternoon and evening;
- values in variable orderHour replaced by three new: divided into three new: morning, afternoon and evening;
- Monday/ Tuesday/ Wednesday/ Thursday/ Friday/ Saturday/ Sunday_couponReceived (values of the day from variable couponReceived transformed to binary variables);
- Monday/Tuesday/Wednesday/Thursday/Friday/Saturday/Sunday_orderTime (values of the day from variable orderTime transformed to binary variables).

4.4. Modelling and Evaluation

In the modelling phase, we applied different algorithms on the prepared data samples. Firstly, we used the original set of attributes as the input dataset and three target attributes. Secondly, for comparison, we decided to use all attributes created in the data preparation phase.

In the initial experiment on the original data, we got quite a high accuracy, ranged from 70-87% for all three target attributes. It was highest for the model with the target attribute coupon3Used (87.3%) using SVM method and coupon3Used using the decision tree algorithm C50 and C45. In the second experiment with the modified data, the results improved in each case. The accuracy ranged from 80-91.5%. The highest accuracy was for coupon3Used (91,5%) using SVM method, coupon3Used (91,3%) using algorithm C5.0. After evaluating all metrics, the algorithms C5.0 and SVM method were ranked among the most successful machine learning methods in solving this task. The least successful method was Naive Bayes, which achieved quite good results for the coupon1Used target attribute, but poorly for the target coupon2used and coupon3Used target attributes, only about 25%.

The following tables present individual the results obtained on the original and prepared data samples.

Table 1 C5.0 algorithm

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	80.7 %	0.62	0.03	0.99	536	123	6	4
c.2	84.3 %	0.62	0	1	564	105	0	0
c.3	87.1 %	0.76	0	1	583	86	0	0

Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	86.1 %	0.93	0.43	0.96	522	73	20	54
c.2	90.0 %	0.93	0.53	0.97	546	49	18	56
c.3	91.3 %	0.93	0.59	0.96	560	35	23	51

We can see, that model was not able to predict class 1 for coupon2 and coupon3 on original data samples. On the other hand, the model was successful for each target attributes on the prepared data samples, and accuracy has improved in each case.

Table 2 C4.5 algorithm

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	79.5 %	0.624	0.09	0.96	520	115	22	12
c.2	84.3 %	0.624	0	1	564	105	0	0
c.3	87.1 %	0.624	0	1	583	86	0	0
Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	85.1 %	0.930	0.65	0.90	487	45	55	82
c.2	87.0 %	0.930	0.50	0.94	530	53	34	52
c.3	87.6 %	0.930	0.45	0.94	547	47	36	39

A similar situation occurred for the C4.5 algorithm. The highest accuracy value was for the coupon 3. The number of incorrectly classified values was higher than within the algorithm C5.0.

Table 3 Support Vector Machine method

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	80.9 %	0.612	0.03	0.99	537	123	5	4
c.2	84.3 %	0.624	0	1	564	105	0	0
c.3	87.3 %	0.600	0.01	1	583	85	0	1
Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	86.1 %	0.94	0.45	0.96	520	70	22	57
c.2	89.3 %	0.94	0.50	0.97	546	52	18	53
c.3	91.5 %	0.94	0.59	0.96	561	35	22	51

The SVM method had excellent results on the prepared data samples. The AUC value was, in any case, equal 94. This method generated one of the most successful models, achieving an accuracy of 91.5%.

Table 4 Logistic regression

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	81.0 %	0.625	0.05	0.09	525	110	17	17
c.2	84.3 %	0.655	0	1	564	105	0	1
c.3	87.2 %	0.624	0	1	583	86	0	1
Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	84.9 %	0.655	0	1	568	101	0	1
c.2	89.8 %	0.855	0	1	601	68	0	1
c.3	90.7 %	0.893	0.056	0.86	590	50	17	12

Logistic regression had the worst problem with a prediction of class 1 in four cases of all six.

Table 5 Random forest algorithm

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	80.8 %	0.624	0.09	0.98	529	115	13	12
c.2	83.9 %	0.600	0.07	0.98	554	98	10	7
c.3	87.0 %	0.556	0.07	0.99	576	80	7	6
Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	85.5 %	0.930	0.65	0.91	492	45	50	82
c.2	89.7 %	0.935	0.50	0.97	546	52	18	53
c.3	90.3 %	0.940	0.40	0.98	570	52	13	34

Random forest, unlike the previous four cases, was able to classify both classes on original and prepared data samples. The model's accuracy has been improved on prepared data. Accuracy and AUC values were the highest for coupon3.

Table 6 Naive Bayes method

Original data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	70.6 %	0.530	0.26	0.81	439	94	103	33
c.2	23.6 %	0.593	0.92	0.11	61	8	503	97
c.3	27.7%	0.534	0.93	0.2	114	15	469	71
Prepared data								
	acc	AUC	sens	spec	Confusion matrix			
					TP	FP	FN	TN
c.1	81.2 %	0.900	0.94	0.78	423	7	119	120
c.2	80.6 %	0.906	0.96	0.78	438	4	126	101
c.3	75.5 %	0.862	0.91	0.73	427	8	156	78

Although NB was able to classify both classes, the accuracy results were very low.

5. CONCLUSIONS

Our work aimed to investigate a performance assessment of different classification methods for digital coupon marketing. We used the CRISP-DM methodology for analyses of the historical data representing the coupon redemption for three possible products in one order. For this purpose, we used two data samples, one original from the Data Mining Cup and the second one as the result of the data preparation phase. In the modelling and evaluation, we applied and compared following machine learning algorithms: C4.5, C5.0, Random Forest, Naive Bayes, Support Vector Machine, and Logistic Regression. Finally, we can say that the C5.0 and Support Vector Machine algorithms are appropriate for the analysis of data on customer buying behaviour. The figure 1 shows the ROC curve of the most successful model created by Support Vector Machine, with a value of specificity 0,96 and sensitivity 0,59. In similar works, the most successful algorithms were XGBoost, RF, and CART.

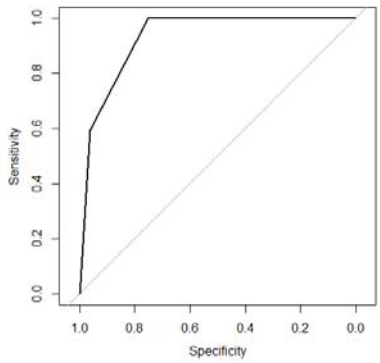

Fig. 1 ROC curve of the most successful model.

REFERENCES

[1] DASKALOVA, N. - BENTLEY, F. - ANDALIBI, N.: It's All About Coupons: Exploring Coupon Use Behaviors in Email. Proceeding CHI EA '17 Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 2017, s.1152-1160.

[2] HE, H. - JIANG, W.: Understanding users´ coupon behaviors in e-commerce environments. In 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications.2017, s.1047-1053.

[3] WU, J. - ZHANG Y. - WANG J.: Research on Usage Prediction Methods for O2O Coupons. In: Cheng L., Leung A., Ozawa S. (eds) Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science. 2018, Volume 11305.

[4] PUSZTOVÁ, Ľ. - BABIČ, F.: Analysis of Users Buying Behaviour to Improve the Coupon Marketing. Business information systems workshops. *BIS*. 2017, s.

[5] CHAPMAN, P. - CLINTON, J. - KERBER, R. - KHABAZA, T. - REINARTZ, T. - SHEARER, C. R. - WIRTH, R.: CRISP-DM 1.0: Step-by-step data mining guide.2000.

[6] SIREGAR, B. - NABABAN, E. B. - SAGALA, N. - ANDAYANI, U.: Tuition Single Classification using Decision Tree Method and C4.5, In Journal of Physics.2019, Volume 1175.

[7] JORDA, E. R. - RAQUENO, A. R.: Predictive model for the academic performance of the engineering students using CHAID and C5.0 algorithm. In International Journal of Engineering Research and Technology. 2019, Volume12 (6), s.917-928.

[8] LAVANYA, B.: Performance analysis of decision tree algorithms on mushroom dataset. In International Journal for Research in Applied Science & Engineering Technology (IJRASET).2017, Volume 5, s. 183-191.

[9] AGRAWAL, G. L. - GUPTA, H.: Optimalization of C4.5 decision tree algorithm for data mining application. In International Journal of Emerging Technology and Advanced Engineering.2 013, Volume 3, s. 341-345.

[10] MIENYE, D. I. - WANG, Z. - SUN, Y.: Prediction performance of improved decision tree-based algorithms: a review. Conference: 2nd International Conference on Sustainable Materials Processing and Manufacturing.2019, s. 698-703.

[11] BREIMAN, L.: Random Forests. Machine Learning.2001, Volume 45, s.5-32.

[12] ALZUBAIDI, L. - ARKAH, Z.M. - HASAN, R. I.: Using Random Forest Algorithm for Clustering. In Journal of Engineering and Applied Sciences.2018, Volume 13, s.9189-9193.

[13] PARALIČ, J.: Objavovanie znalostí v databázach, Košice.2003.

[14] BERRAR, D. - LOPES, P. - DUBITZKY, W.: Incorporating domain knowledge in machine learning for soccer outcome prediction.

[15] EZUKWOKE, K. - ZAREIAN, S.: Logistic regression and kermel logistic regression: A comparative study of logistic regression and kernel logistic regression for binary classification. 2019.

[16] MUNOZ, A. - MOGUERZA, J. M. - MARTOS, G.: Support Vector Machines. Wiley StatsRef: Statistics reference online. NCH Marketing Services.2019.

[17] DATA MINING CUP, https://www.data-mining-cup.com/reviews/dmc-2015/

Received April 17, 2020 , accepted May 13, 2020

BIOGRAPHIES

Ľudmila Pusztová was born on 16. 02. 1993. In 2017 she graduated (MSc) at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. Since September 2017 she works as PhD. student at the Department of Cybernetics and Artificial Intelligence. The dissertation for PhD study is focused on models and methods of data analysis for the creation of knowledge models from data sources. Currently, she is working on resolve the most critical problem in the case-based reasoning method - adaptation, which is often done manually by the experts in the relevant field.

František Babič graduated (MSc.) at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice in 2005. In the same year, he began his doctoral study in the same department and successfully finished in 2009. In 2018 he became an associate professor. He has been participating in several international and national research projects, such as FP6 IST, COST, Central Europe, APVV, VEGA, KEGA. His scientific research is focusing on data analytics, knowledge processes, and project management. He is the author of over 150 scientific publications (more than 40 indexed in the WoS database).