A Churn Analysis Using Data Mining Techniques: Case of Electricity Distribution Company

Jiri Pribil, *Member, IAENG*, Michaela Polejova

***Abstract*— This paper focuses on issues of a prediction of the probability of a customer leaving for competition. The cost of acquiring a new customer is typically several times higher than the cost of retaining a current customer. Churn modeling is a powerful tool to help target retention activities more accurately. A real dataset with customer data over which a churn model is created using logistic regression and the decision tree is used in this paper. The CRISP-DM methodology is applied to the entire process. Based on a critical assessment of the modeling process and its outputs, recommendations are given for further work with the models and for improving their quality.**

***Index Terms*—CRISP-DM, decision tree, churn, data mining, logistic regression**

I. Introduction

CUSTOMER retention has become the central theme of marketing for both foreign and domestic firms in recent years. After 2006, the energy market was opened in the Czech Republic and the following years a strong competitive struggle developed among existing and newly established traders. Customers have begun to change their vendors in huge amount. This phenomenon is commonly known as churn.

The loss of the customer means the loss of any future revenue for the company. The cost of acquiring a new customer is up to ten times higher than the cost of maintaining an existing customer [1]. At the same time, however, retention activities cannot focus on the entire customer portfolio or only on randomly selected customers or groups of customers. The solution for this problem can be modeling of customer loss or so-called churn modeling. The predictive churn model can predict the likelihood of current customers leaving based on historical data of leavings.

The reasons for churn differ slightly depending on the industry. [2] provide the most common factors determining customer churn in service provision: service price, service provider change, a competitor with better technology, quality of service, customer satisfaction with service, security of personal data, and service advertising.

Manuscript received July 23, 2017; revised August 11, 2017.

Jiri Pribil is an Assistant Professor and Vice Dean for Studies at the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (e-mail: jiri.pribil@vse.cz).

Michaela Polejova is a student at the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (e-mail: xpolm62@fm.vse.cz).

Churn modeling falls into the data mining area. Data mining applications are currently projected into many areas. Banks predict the risk of credit default, insurance companies detect fraud in the payment of benefits and reimbursements, direct marketing tries to target a promising customer, forensic science estimates the location of a future crime, schools predict the likelihood of the student terminating the study prematurely. In the healthcare, data mining is used to determine the diagnosis and determination of the treatment method, etc. [3].

II. Data mining

A. Data Mining Process

Data mining was originally part of the process of knowledge discovery in databases (KDD) using application-specific algorithms for extracting patterns from data. Over time, however, the concept of data mining has expanded and this has become synonymous with the knowledge discovery [4]. In contemporary literature, there is little to distinguish between these two concepts.

[5] describes the knowledge discovery process as a sequence of five phases, with the data at the beginning and the knowledge at the end – (a) selection, (b) preprocessing, (c) transformation, (d) data mining, and (e) interpretation/evaluation. At the same time, he adds that the process includes some additional steps, such as understanding of the problem, setting goals, and utilizing the outputs (knowledge) in practice.

The process of data mining describes [4] as follows: (a) definition of the objectives for analysis, (b) selection, organization and pre-treatment of the data, (c) exploratory analysis of the data and subsequent transformation, (d) specification of the statistical methods to be used in the analysis phase, (e) analysis of the data based on the chosen methods, (f) evaluation and comparison of the methods used and the choice of the final model for analysis, and (g) interpretation of the chosen model and its subsequent use in decision processes.

By comparing the contents of the individual phases of both processes, it can be determined that the data mining process, as is known today, is equivalent to the knowledge discovery in databases process when considering the above-mentioned additional steps in the KDD process. Other authors writing about data mining are almost identical in the description of the data mining process. [6] lists the following stages: (a) define goal, (b) select data, (c) prepare

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

data, (d) select and transform variables, (e) process model, (f) validate model, and (g) implement model.

Most authors refer to the CRISP-DM methodology in their works, which will be discussed in the next chapter.

B. CRISP-DM Methodology

CRISP-DM methodology (Cross Industry Standard Process for Data Mining) began at the end of 1996. The data mining market was still young at the time, but the interest rate rose rapidly. Individual companies have developed their own approaches and methodologies through trial and error. There was a need to develop a unified and standardized process model based on experiences and freely available to all those interested in data mining in business processes in different industries. The methodology has been developed and refined over several years so that in 2000 it could be published and disseminated among the public.

The methodology is expressed through a hierarchical process model. The model consists of a series of tasks that are divided into four levels, from general to specific. At the highest level, the data mining project consists of phases. Each phase also consists of general tasks that must be applicable to all possible data mining situations and types of projects. The third level is a specialized task that puts the tasks from the second level in different specific situations and concretizes them. On the fourth level, there are procedural examples – records of specific actions, decisions and results that describe what happened in a specific data mining project.

The CRISP-DM process model [7] provides a general overview of the life cycle of the data mining project consisting of six phases and their usual continuity. However, these are not fixed, it is almost always necessary to return to the previous stages and to reassess the individual steps and decisions. The process of data mining itself has a cyclical nature – what the investigator can learn during the entire process can lead him to new, better, more relevant, or more critical issues to be addressed, and he finds himself again at the very beginning of the process.

III. Understanding the Problem

Following the full opening of the electricity and gas market in the Czech Republic in 2006, energy companies have begun to face the problem of leaving customers who could freely choose their electricity and gas suppliers [8]. In addition to traditional suppliers, alternative suppliers – emerging companies dealing with one or both commodities, which do not own any part of the distribution network – have begun to appear.

In 2009, the customer churn came to such high numbers that it was necessary to start actively addressing the situation. Energy companies responded to the situation by creating sales channels through which they wanted to attract new customers and more retention products that have been actively offered in retention campaigns.

One way to work with outbound customers is to use *reactive retention*. In this process, the customer is approached when he has already filed the termination of the contract. The notice period ranges from about 20 days (fixed-term products) to 3 months (products of unlimited duration). During this time, the customer is offered a product with a better price, or other benefits. If the customer agrees to the offer, a new contract is concluded with him. The undisputed advantage of this process is the fact that the company knows what new supplier the customer wants to join, so he can offer him a better price than the competitor. However, the disadvantage is the complexity of the process – communication with the customer, sending and completing documents, knowledge of the deadlines, and the need to respond quickly during the changeover process.

Proactive retention, unlike reactive, tries to reach out to the customer before submitting the contract termination. Ideally, a company with predictive modeling knows that the customer will want to leave before he knows it.

A. Determining Business and Data Mining Objectives and Success Criteria

The primary business objective of the project is to keep current household customers taking electricity by predicting the probability of their leaving. There are many other commercial issues that are related to the problem:

1) What are the criteria for deciding who to try to keep as a customer?
2) Do we include the customer value criterion in decision making?
3) If so, how do we set this value?
4) When to contact the customer (how long before the end of the contract)?
5) How do we reach the customers we want to keep, what we should offer them?
6) How to deal with a situation where the customer will negotiate with us?
7) How to balance the budget with the number of customers we want to reach?

The expected commercial benefit is to reduce the exit rate with the consequent effect of lowering the margin squeeze. The number of retained customers, respectively the reduction in churn rate will be considered as the success criterion.

The primary data mining goal is to create a prediction model identifying customers prone to leaving the company. The secondary goal of data mining is to identify the variables that have the highest impact on customer attrition rate. The criterion of data mining success is the quality of the models that will be created, namely the precision of prediction of the lost customers, the model's ability to correctly classify lost customers, and the overall prediction accuracy.

B. Assessment of the Situation

At present, the company uses the KXEN software, which has a simple and user-friendly environment over the backend MySQL database (KXEN DB), which is linked to SAP BW (SAP Business Warehouse) data warehouse, which gathers data from enterprise information systems, especially SAP IS-U and SAP CRM. Fig. 1 illustrates the situation.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

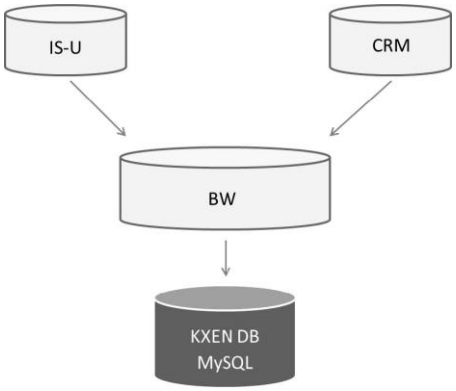

Fig. 1. Overview of existing data sources.

IV. Data Understanding

The data sample to be used for modeling comes from the KXEN summary table. There is also a reference table for the assignment of codes of municipalities from corporate system to codes used by the Czech Statistical Office (CZSO), the CZSO data file with data on the number of inhabitants in municipalities and the CZSO data file with data on population in regions and districts in the Czech Republic. In addition to the numbers of inhabitants, the average age is shown in these tables.

The CZSO data on the population and the average age will be attached to the main dataset for the sake of clarity and for the purposes of subsequent steps.

A. Description and Exploration of Data

The data file contains 30,000 rows (records) and 96 columns (attributes, variables). It contains a data sample of mass-served household customers who, in 2012, had an active contract or, in the same period, left the monitored company for another electricity supplier. The ratio of active to lost customers is 80:20 (24,000:6,000 customers). The sample includes customers who have entered into a contract with the company by the end of 2010. Customers are billed once a year, and consumption and billing data refer to 2010-2011.

Based on knowledge of the content of variables, it is now possible to determine which of them will be excluded from the subsequent. This initial exclusion of some variables will simplify further work with the dataset. A total of twenty variables, which are not relevant to the required analysis, were excluded from further analysis. For example, the household/business customer category (all data are related to households), data on gas supplies, data on the legal form of the organization, etc.

After the initial exclusion of variables, 76 attributes were left in the dataset.

B. Qualitative Variables

The model presents both the qualitative variables and categories that variables can take. Within the survey of these variables, the differences in the impact of the categories on the target variable LOST (e.g. churn rate of company customers) will be monitored.

Fig. 2 shows an example for BP_GENDER variable (gender of the customer).

		BP_GENDER		Retained Customers		Lost Customers	
Category	Note	AF	RF [%]	AD	RD [%]	AD	RD [%]
0	missing value	557	1.86	437	78.46	120	21.54
1	man	16,178	53.93	13,507	83.49	2,671	16.51
2	woman	13,265	44.22	10,056	75.81	3,209	24.19
Total		30,000	100.00	24,000		6,000	

Fig. 2. Table of frequencies and distribution of the target variable BP_GENDER in the categories. AF/RF – absolute/relative frequency, AD/RD – absolute/relative distribution.

C. Quantitative Variables

Basic statistics were calculated for numeric variables. The most customers have a permanent residence at the power consumption site, so the variables on population and the average age in municipalities, districts, and regions of the customer's permanent residence will be excluded and only a consumption site will be monitored.

In addition to basic statistics, charts were constructed for numerical variables – Fig. 3 shows an example for BP_AGE variable (age of customer).

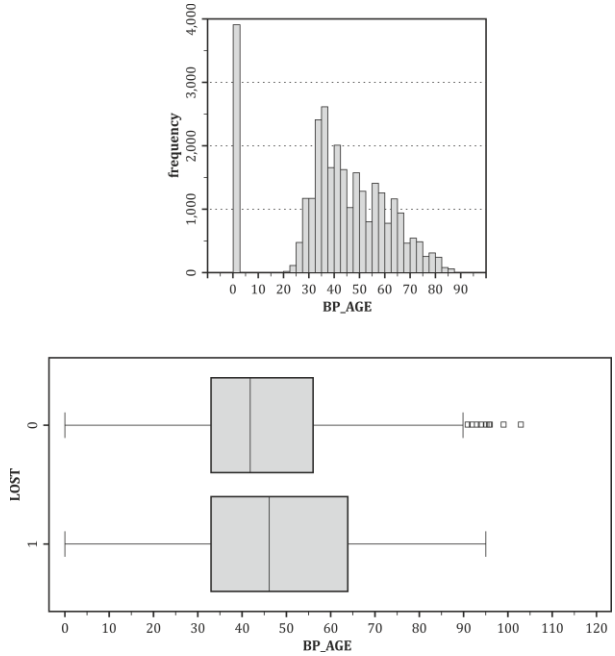

Fig. 3. Histogram and box graph of BP_AGE variable.

D. Data Construction

Several new variables were created in the data set and the original variables used to derive the new ones were removed. These variables include, for example:

1) Binary variable for one or more supply points.
2) Binary variable for 0 or at least one gas supply point.
3) A variable expressing the length of the customer's contract, calculated from the current date and date of signature of the contract.
4) The number of changes in the contract.
5) A variable reflecting the change in consumption over the reference period.

The final data set contains 37 attributes and 30,000 rows.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. Modelling

For modeling of customer churn, a logistic regression was chosen with the target attribute LOST, which distinguishes between lost (1) and retained (0) customers. The *RapidMiner* software used for logistic regression does not use the classical logit model, but the support vector algorithm. There are some differences from the classic logistic regression – the *Logistic Regression* operator requires a dataset with numeric attributes and a nominal target variable. The *Nominal to Numerical* operator has been used to convert the nominal attributes to numeric, and the attributes have been converted to integers.

The decision tree was constructed using the *Decision Tree* operator to compare the methods and to understand the characteristics of lost customers.

The performance of the models has been validated by cross validation which helps to evaluate how the model will behave on new data. After the cross validation (with 3, 5, and 10 validations) on several models, it was found that the models behave best with 10 validations. This number is generally referred to as optimal in the literature [9].

To debug the models, parameter *C* was used (default set to 1.0), which specifies a tolerance for poorly classified cases – a higher value means a freer boundary, a lower tight boundary.

A. Model A

In the first phase, the simplest procedure was tried - all variables were included in the model and subsequently, the parameter *C* was changed. The model showed the most satisfactory result C = 1.0. Fig. 4 shows the confusion matrix and ROC curve of this model. The accuracy of predicted churn 30.61%, 44.48% of the actually lost customers was predicted correctly. The overall accuracy of the model reached 68.47%, the Area Under Curve (UAS) is 0.716.

	true 0	true 1	class precision [%]
predicted 0	4917	925	84,17
predicted 1	1680	741	30,61
class recall [%]	74,53	44,48	

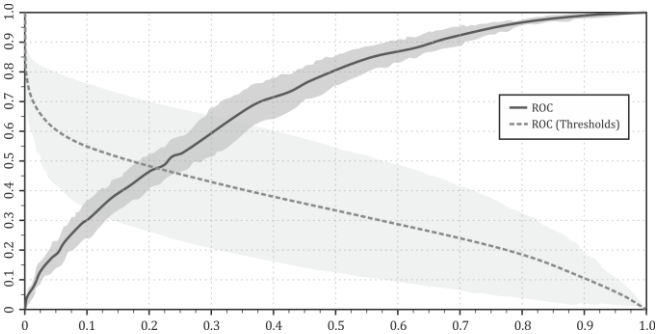

Fig. 4. Model A – Confusion matrix and ROC curve.

B. Model B

In the next step, the number of variables was corrected by removing the correlated variables through the *Remove Correlated Attribute* operator, selecting the variables according to their weights calculated by the χ^2-test performed by the *Weight By Chi Squared* operator, and the *Information Gain Ratio* operator.

With the decreasing number of variables according to the χ^2-test and the *Information Gain Ratio*, the overall accuracy of the model almost did not change, but the recall rate declined.

The best model is represented by the substitution matrix and the ROC curve – see Fig. 5. The prediction of lost customers is 40.21% accurate. The truly lost customers were then correctly classified at 29.59%. The total rate of the correctly classified customers was 76.93%. The area under the ROC curve is 0.715.

	true 0	true 1	class precision [%]
predicted 0	5864	1173	83,33
predicted 1	733	493	40,21
class recall [%]	88,89	29,59	

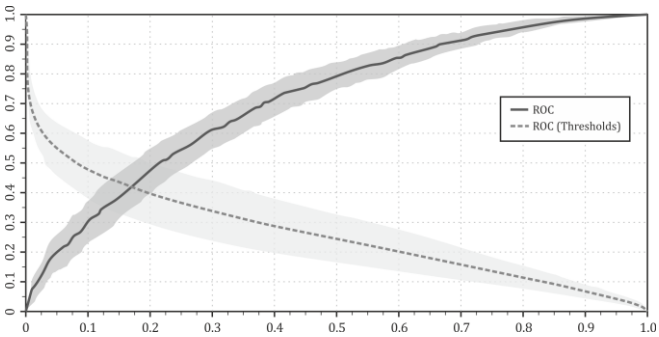

Fig. 5. Model B – Confusion matrix and ROC curve.

C. Model C

Involving *Forward Selection* and *Backward Elimination* operators brought only worsening to the model, especially for the recall rates. Satisfactory results were achieved using the *MRMR (Minimum Redundancy Maximum Relevance)* operator, which iteratively adds attributes with the highest information value relative to the target variable and the least redundancy relative to the attributes already selected.

Fig. 6 shows the confusion matrix that eliminates two attributes through the *Remove Correlated Attributes* operator (*Expected electricity consumption* and *Total average annual invoice on the contract*) and four attributes through the *MRMR* operator (*Indicator of the registered phone number*, *Capacity of the circuit breaker*, *Population of the municipality*, *Combined billing*).

The accuracy of the predicted lost customers is 42.92%. Of the total 1,666 lost customers, the model correctly classified 23.29%. Fig. 6 shows the confusion matrix and ROC curve of this model. The area under the curve is 0.751, the overall prediction capability of the model is 78.29%.

	true 0	true 1	class precision [%]
predicted 0	6081	1278	82,63
predicted 1	516	388	42,92
class recall [%]	92,18	23,29	

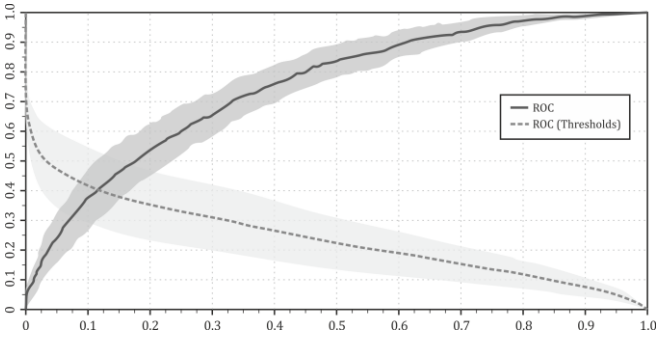

Fig. 6. Model C – Confusion matrix and ROC curve.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

D. Model D

Various combinations of parameters have been tested in the decision tree. The best results were obtained using the *Gini index* criterion. Fig. 7 shows that the prediction accuracy of lost customers in the decision tree reached almost 81%. The model has been able to classify 53.48% of the truly lost customers. The overall number of the correctly classified customers was 88%. The area under the curve is 0.810.

	true 0	true 1	class precision [%]
predicted 0	6387	775	89,18
predicted 1	210	891	80,93
class recall [%]	96,82	53,48	

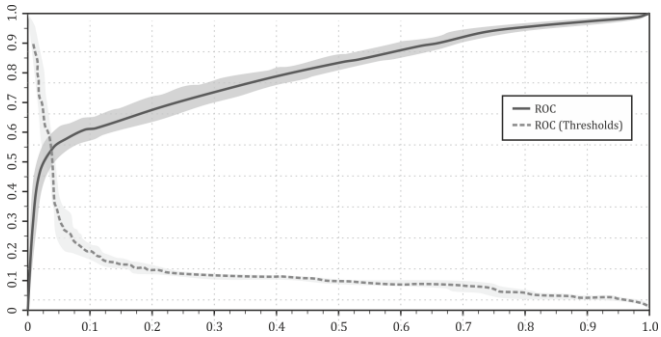

Fig. 7. Model D – Confusion matrix and ROC curve.

The decision tree answers the question of determining the most important variables. It is also possible to derive decision-making rules from it.

E. Models Comparison

Fig. 8 provides a summary of the performance characteristics of each model. The most accurate model is Model D – the decision tree, which has reached the highest values in all characteristics.

Among the models created by the logistic regression, Model C has the best values in three categories: accuracy, AUC, and precision, but has the lowest ability to correctly classify.

When deciding on a used model, it is necessary to consider the costs of the activities that will be commenced during the churn management. In practice, the situation is likely to be more expensive if, the effectively lost customer is classified as retained. These customers will not be cared during the churn process, so they will go further and take their money with them. In a situation where the model predicts the active customer as lost, there will be unnecessary costs for applying some churn steps, but it is likely to be a smaller loss than in the previous situation.

Criterion	Note	Model A	Model B	Model C	Model D
Accuracy [%]	overall accuracy of the model prediction	68.47	76.93	78.29	88.08
AUC	Area under the ROC curve	0.72	0.72	0.75	0.81
Precision (1) [%]	accuracy of predicted lost customers	30.61	40.21	42.92	81.02
Recall (1) [%]	ability to correctly classify the lost cust.	44.48	29.59	23.29	53.48
F-measure (1) [%]	harmonic average precision and recall	36.30	34.10	30.20	64.40
Kappa	degree of consent (actual vs. forecast)	0.16	0.21	0.19	0.58

Fig. 8. Comparison of models A-D.

VI. Conclusions

The main objective of this project was to create a functional prediction model identifying customers that are vulnerable to leaving the company. In the modeling phase, 46 models were created, the best of which were compared to each other. The most accurate model was a decision – the precision of 88% should be considered to be very high.

The secondary goal of this project was to identify the main variables that affect customer churn. The most influential variables were (a) the consumption change category between two billing cycles, (b) product type, (c) high tariff consumption, (d) total consumption, (e) estimated consumption, (f) age, (g) the month of the billing cycle, (g) contract length, (h) low tariff consumption, and (i) monthly payments.

A. Churn Management

Churn modeling should be seen as part of higher units – churn management, customer lifecycle management, customer relationship management. The basis for all these activities is customer data, correct, complete, up-to-date data. To exploit the data potential for modeling and subsequent decision making, it is necessary to provide one good analytical data market.

B. Proposal to Improve Data Quality

The data base needs to be maintained in the best possible condition. For this purpose, it is possible to use both inbound and outbound customer calls with the call center, where the operator can see customer data in the CRM system and has the possibility to verify or add some of them. Newly concluded customer contracts provide another opportunity, whether for a commodity or for additional service. Marketing research can also be a source of data for customers to understand their attitudes and preferences

C. Application of Ensemble Modelling

Ensemble modeling is a data mining area that deals with model file creation. In the case of a company, the customer database could be randomly divided into several parts, a classification churn model would be created above each part, and these models would then be combined. In addition to completely random distribution, the above-mentioned breakdown of customer segments could be used. Combining models means, in most of the cases, a significant increase in prediction ability. Ensemble modeling can combine tens and hundreds of models. The disadvantage is more complex interpretations when it is not always clear which factors have contributed to increasing predictive performance. The best-known techniques of combining models are bagging, boosting and stacking [9].

D. External Data

It would be useful to use some external data as an addition to the customer data – e.g. municipal statistics, data from direct marketing agencies, etc.

E. New Variables

At the data exploration phase, it was found that many of the variables are duplicated and some variables have 90% of all records in only one category. Based on the conclusions of

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

this phase, new variables were proposed to help improve the quality of the models: (a) customer value (CLV), (b) customer-specific margins, (c) acquisition cost per customer and the channel through which the customer was acquired, (d) end date of the contract (for fixed-term products), (e) year-on-year changes in consumption and invoice history for the entire duration of the contract, (f) data about the online customer portal (frequency of sign-up, changes made), and (g) data about the contacts with customer center (inbound vs. outbound calls, e-mail) for example in the last 12 months.

The presented work is a partial contribution to data mining and its application for a commercial area. There is room for further elaboration of the discussed topic both in the methodical width and in the application depth.

REFERENCES

[1] G. Olle, "A Hybrid Churn Prediction Model in Mobile Telecommunication Industry," *Int. J. E-Educ. E-Bus. E-Manag. E-Learn.*, 2014.
[2] R. Hejazinia and M. Kazemi, "Prioritizing factors influencing customer churn," *Interdiscip. J. Contemp. Res. Bus.*, vol. 5, no. 12, pp. 229--236.
[3] E. Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley & Sons, 2013.
[4] P. Giudici, *Applied Data Mining: Statistical Methods for Business and Industry*. John Wiley & Sons, 2005.
[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, p. 37, Mar. 1996.
[6] O. P. Rud, *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*. John Wiley & Sons, 2001.
[7] P. Chapman *et al.*, "CRISP-DM 1.0." SPSS, 2000.
[8] M. Polejová, *A Churn Analysis Using Data Mining Techniques*. 2015.
[9] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)