Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia

Biljana Stojkoska, Aleksandar Petkovski

Analysis of Churn Prediction: A Case Study on Telecommunication Services in Macedonia

Aleksandar J. Petkovski, Biljana L. Risteska Stojkoska,
Kire V. Trivodaliev, and Slobodan A. Kalajdziski

***Abstract* — Customer churn is one of the main problems in the telecommunications industry. Several studies have shown that attracting new customers is much more expensive than retaining existing ones. Therefore, companies are focusing on developing accurate and reliable predictive models to identify potential customers that will churn in the near future. The aim of this paper is investigating the main reasons for churn in telecommunication sector in Macedonia. The proposed methodology for analysis of churn prediction covers several phases: understanding the business; selection, analysis and data processing; implementing various algorithms for classification; evaluation of the classifiers and choosing the best one for prediction. The obtained results for the data from a telecommunication company in Macedonia, should be of great value for management and marketing departments of other telecommunication companies in the country and wider.**

***Keywords* — churn prediction, data mining, decision trees, KNN, logistic regression, naïve Bayes.**

I. Introduction

Customer churn is an important issue that is often associated with the life cycle of the industry. When the industry is in a growth phase of its life cycle, sales are increasing exponentially and the number of new customers largely outnumbers the number of churners. On the other side, companies in a mature phase of in their life cycle, set their focus on reducing the rate of customer churn [1].

The main reasons that cause customer churn are divided into two groups: accidental and intentional. Accidental churn happens when the circumstances are changing so prevents the customers from using the services in the future. Examples of accidental churn are economic circumstances that make services too expensive for the customer. Intentional churn occurs when the customers choose to switch to another company that provides similar services. This type of churn is the one that most companies are trying to prevent. An example of intentional churn are better offers from competition, more advanced services and better price for the same service [2].

Establishing a system for managing the customer churn is vital. There are two basic approaches for managing customer churn: directed and undirected. In undirected approach, companies rely on superior product and mass advertising to increase loyalty to the brand and to retain customers. In direct approach, companies rely on identifying customers who are likely to churn, and then to adapt their requirements to prevent from churning [3].

In the recent years, churn prediction is becoming very important issue in the telecommunications industry [4][5]. In order to deal with this problem, the telecom operators must recognize these customers before they churn. Therefore, developing a unique classifier that will predict future churns is vital. This classifier must be able to recognize users who have a tendency to churn in the near future, so the operator will be able to react promptly with appropriate discounts and promotions. The most frequently used techniques for this purpose are learning algorithms for classification, like decision trees, logistics regression, *k*-nearest neighbors, Naïve Bayes, neural networks, etc. [6][7]. Moreover, researches should focus on identifying new features that are most effective in predicting the customer churn [8].

In this paper, we aimed to investigate the main reasons for churn among fixed-telephony subscribers in Macedonia. For this purpose, we gathered and processed the data, and based on these data, we implemented and compared four well known machine learning algorithms. Additionally, we identified the most important factors which are crucial for the customers to churn, that are tariff plan, subscriber contract, duration (length) of the contract, number of services, number of outgoing calls per month, and average call duration in the last month.

The rest of this paper is organized as follows: The next section defines the methodology for churn prediction, regarding different phases of the process. It also describes the algorithms used for churn prediction. Section III presents the results of the classification algorithms applied on fixed telephony subscribers from one Macedonian provider. This paper is concluded in Section IV.

II. Methodology for Churn Prediction

In order to find a possible solution to the problem of churn prediction i.e. successfully apply a machine learning technique to the available data, one needs a deep

Aleksandar Petkovski is with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia; (e-mail: petkovski.aleksandar@yahoo.com).

Biljana Risteska Stojkoska is with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia; (e-mail: biljana.stojkoska@finki.ukim.mk).

Kire Trivodaliev is with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia; (e-mail: kire.trivodaliev@finki.ukim.mk).

Slobodan Kalajdziski is with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia; (e-mail: slobodan.kalajdziski@finki.ukim.mk).

978-1-5090-4086-5/16/$31.00 ©2016 IEEE

understanding of the business rules of the telecommunications company and their specificity. Such knowledge enables the selection of attributes suitable for the problem at hand. The quality of the data can further be improved by subjecting it to preprocessing. Once a final dataset is derived, the classification algorithms can be successfully trained and their performances correspondingly evaluated. In the following subsections, we present the identified phases in our methodology.

A. Business Understanding

In this initial phase, the focus is set on understanding the project objectives and requirements from the telecommunications business perspective. The aim of the churn prediction is to identify the properties that make a customer churn in order to prevent it and retain the customer. To enable this, we consider customers that churned and analyze their data over a period while they still used the services of the telecommunications company.

B. Data Understanding

For the purpose of this paper, a telecommunication company from Macedonia shared their data, through text files exported from certain tables in their database. We have anonymized the data (we only care about the user's dynamics data, not their personal data). The obtained data covers 28 months period from 01.01.2012 to 30.04.2014 (approximately 34 million records). Additional data for the customer complaints is included in the dataset since it is a strong indicator for customer dissatisfaction.

C. Data Pre-processing

The data pre-processing tasks include careful selection of data attributes and records. Because we deal with incomplete and noisy data, some additional data cleaning and transformation are also performed.

1) Data Selection

We first identify and extract the most relevant attributes for the research. The initial dataset consists of 68 (mainly numeric) attributes, that can be grouped in the following three categories:

- **Demographic attributes:** contain the primary features of the customer such as sex, age, nationality, place of residence, etc.
- **Contract attributes:** contain the attributes associated with the customer contract for a particular service such as type of service, date of conclusion of the contract, price of the service etc.
- **Customer behavior attributes:** describe the customer activities.

The data subjected to our analysis spans over a period of one year from 01.05.2013 to 30.04.2014. A total of 22461 customers are included, of which 2629 customers are churns, while 19832 customers still use the services of the telecommunication company.

2) Data Cleaning

The presence of noise, unknown or empty values, outliers and invalid values may negatively affect the performance of the machine learning algorithm by using the raw data. The purpose of data cleaning is to reduce the number of inconsistent values, remove noise and incomplete entries and attributes. Since our dataset is sufficiently big, we removed all potentially problematic tuples.

3) Data Transformation

Data transformation techniques can significantly improve the overall performance of the churn prediction, which we have seen while experimenting with potential transformations. The prediction produces best results when data attributes are normalized (in the [0,1] range) and discretized (Sturges and *k*-proportional were used) and the results presented in this paper refer to such data.

4) Feature Selection

Features selection refers to the process of selecting a subset of relevant attributes of a set of attributes. This reduces the number of input attributes to the learning algorithm, thereby significantly reducing time and resources required to train the algorithm. For feature selection, we have chosen the following techniques:

- **Chi-Squared:** Based on statistical theory, chi-squared evaluates the attributes values by calculating the value of the chi-squared statistics on each attribute regarding the class. The formula for calculating the value is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_I)^2}{E_i} \qquad (1)$$

  where O_i is the observed value, E_i is the expected value and n is the number of rows.

- **Information Gain:** Algorithm that evaluates attributes by measuring the information gain regarding the class. Information gain is obtained through the following formula:

$$Gain(A) = E(S) - \sum_{i=1}^{t} \frac{S_i}{S} Entropy(S_i) \qquad (2)$$

  where $E(S)$ is the entropy of the given set, $Entropy(S_i)$ is the entropy of the i-th subset obtained by the division in terms of attribute A.

We experimented with different feature selection techniques and tested the performances of different classifiers. The feature selection process resulted in two datasets (42 attributes and 17 attributes). The performances are better with the reduced attribute dataset and in this paper the results presented for these classifiers refers to such usage.

D. Machine learning approaches for churn prediction

There are many techniques that have been proposed for customer churn prediction. In our approach, we will analyze four machine learning algorithms: C4.5 decision tree, *k*-nearest neighbors algorithm, naïve Bayes classifier and logistics regression.

1) C4.5

The C4.5 classification algorithm uses the concept of entropy as follows. Suppose that we have a candidate split S, which partitions the training data set T into several subsets, $T_1, T_2, ..., T_k$. The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

$$H_s(T) = \sum_{i=1}^{k} P_i H_s(T_i) \qquad (3)$$

where P_i represents the proportion of records in subset i. We may then define our *information gain* to be gain(S) = $H(T) - H_S(T)$, that is, the increase in information produced by a partitioning the training data *T* according to this candidate split *S*. At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain(*S*).

2) K-nearest neighbors

K-nearest neighbors algorithm compares a test tuple with trained tuples that are similar to it (learning by analogy). The trained tuples are described by *n* attributes (a point in *n*-dimensional space). For a new tuple the algorithm *k*-nearest-neighbors searches the space for *k* trained tuples that are closest to the unknown tuple. Most common class of *k* nearest neighbors is set as class attribute to the new tuple. When *k* = 1, the new tuple is set the same class as the trained tuple which is closest to it in the considered space. The distance between two tuples is calculated in terms of metric distance, such as the Euclidean distance, which is obtained through the following formula:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i}\text{-}x_{2i})^2} \qquad (4)$$

where X_1 and X_2 are the tuples, x_{1i} and x_{2i} are the values for the attributes and *n* is the number of attributes.

3) Naïve Bayes

The Naïve Bayes classifier uses the idea that attributes of the classified objects do not have existential correlation. Based on the training data, the naïve Bayes algorithm calculates the probability of an outcome, given a value of a specific attribute, by taking all attributes to be conditionally independent. The algorithm uses the Bayes' formula to make its predictions:

$$P(c_j|x_i) = \frac{P(c_j)P(x_i|c_j)}{P(x_i)}. \qquad (5)$$

For a new sample $\boldsymbol{x}$ the $P(c_i|\boldsymbol{x})$ should be calculated. If all of the attributes are independent, the probability $P(c_i|\boldsymbol{x})$ is calculated by the following formula:

$$P(c_j|x) = P(c_j)\prod_{i=1}^{n} P(x_i|c_j) \qquad (6)$$

where c_j is the class and x_i is test attribute. The new sample will be set class for the biggest value of probability $P(c_j|x)$.

4) Logistics Regression

Logistic regression is a special case of a linear classifier. When applied to a classification problem, it predicts the class using binary dependent variables instead of continuous. In regression, the dependent variable is the probability of an event to occur. Therefore, the result of applying logistic regression is the relationship between the probability that an event will occur or not. This algorithm uses the logistics function to determine the class of new sample:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad (7)$$

The class value that is closest to $\pi(x)$ will determine the class of the new sample.

E. Evaluation

The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use stratified 10-fold cross-validation. The data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets (each of which have a lot in common). Finally, the 10 error estimates are averaged to yield an overall error estimate. The information of actual and predicted samples is presented using the confusion matrix (Table 1).

TABLE 1: CONFUSION MATRIX.

		Predicted samples	
		True	False
Actual samples	True	TP	FP
	False	FN	TN

Through this confusion matrix, the accuracy of the model can be calculated with the following formula:

$$AC = \frac{TP+TN}{TP+FP+FN+TN} \qquad (8)$$

The error rate is calculated by the following formula:

$$Error\ Rate = \frac{FP+FN}{TP+FP+FN+TN} \qquad (9)$$

AUC (area under a ROC curve) has a value that is in the range between 0 and 1. AUC is an important feature of a model as it presents the probability that the model will rank a randomly chosen positive sample higher than a randomly chosen negative sample. The formula for calculation of the AUC is:

$$AUC = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP\, dFP \qquad (10)$$

where: *P* = *TP* + *FN*, and *N* = *FP* + *TN*.

III. RESULTS AND DISCUSSION

We have tested all proposed machine learning algorithms on the original dataset and reduced attribute datasets, and Table 2 and Table 3 shows their best performances in terms of confusion matrices and optimal number of attributes and their accuracy and area under the ROC curve, respectively.

TABLE 2: CONFUSION MATRICES FOR DIFFERENT CLASSIFIERS.

		Predicted class	
		Non-churner	Churner
Actual class	Non-churner	(KNN) 14440 (C4.5) 14342 (NB) 13329 (LR) 14555	(KNN) 301 (C4.5) 399 (NB) 1412 (LR) 186
	Churner	(KNN) 1292 (C4.5) 1029 (NB) 1088 (LR) 771	(KNN) 908 (C4.5) 1171 (NB) 1112 (LR) 1429

TABLE 3: PERFORMANCES OF DIFFERENT CLASSIFIERS.

Classifier	***#attributes***	***Accuracy***	***AUC***
Naïve Bayes	17	85.243	0.822
C4.5	68	91.571	0.844
k-nearest neighbors	17	90.597	0.857
Logistics regression	68	94.351	0.928

The churn prediction using logistic regression displays highest accuracy of 94,351%, while the Naïve Bayes classifier has the lowest accuracy of 85,243%, which was expected, since the independence assumption in this algorithm is too strong for our data. In terms of the time needed for training, the performances of the machine learning algorithms are similar, with exception of the logistic regression, which is several times slower than the rest, because of its iterative nature.

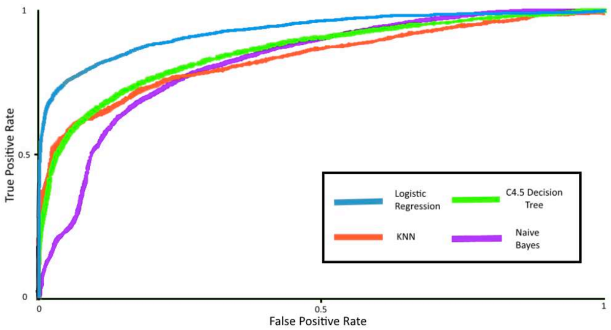

Fig. 1. ROC curves for different classification models.

The classifiers obtained with Naïve Bayes, C4.5 and *k*-nearest neighbors have similar AUC, and it is in the range from 0,822 to 0,857. Once again the best performing algorithm is the logistic regression with AUC of 0,928. In Figure 1 we give the ROC curves for all four classifiers. As can be seen the ROC curve for logistics regression model displays the best threshold for separating samples into appropriate classes.

As can be seen from the comparison of the algorithms, they all produce good results with high accuracy, but it is hard to choose which classifier is the best, since all have both advantages and disadvantages. The classifier obtained by logistic regression shows the best results, but the time for generating and the necessary resources are very large. The generation of this classifier is an expensive solution. The classifiers obtained by C4.5 decision trees and *k*-nearest neighbors show lower performance as compared to logistic regression, however they still have high accuracy. The time needed for the construction of these classifiers is very small. The C4.5 based classifier requires fewer resources than the *k*-nearest neighbors classifier, since the search for the closest neighbors in sets with millions of instances is both time and memory consuming. Additionally, from the C4.5 classifier we could identify the most important factors which are crucial for the customers to churn. The most important attributes are tariff plan, subscriber contract, duration (length) of the contract, number of services, number of outgoing calls per month, and average call duration in the last month.

IV. Conclusion

The telecommunication industry in the recent years is a subject of major changes and from a fast-growing industry has come to a state of saturation accompanied with strong competitive market. Customers starve for better services and prices, while their requirements are extremely complex and difficult to understand. In order to cope with this problem, researches in this area have the following objectives: finding the influential factors for the customer churn, as well as building a classifier for predicting the customer churn.

In this paper we have explained the methodology for building a classifier models that will predict the customer churn from the data from a fixed telephony operator in Macedonia. We have carefully extracted the customers' behavior within a one-year period, resulting in dataset containing 22461 customers. The results from this study show that predicting the customer churn can be successful with high accuracy. The classification models derived from C4.5, *k*-nearest neighbors and logistic regression have an accuracy of over 90%. The highest accuracy is achieved with logistic regression with 94,351% accuracy. The disadvantage of this classifier is its execution time and the need for the vast memory resources. The models based on decision trees also shows high accuracy (91.57%), while Naïve Bayes and *k*-nearest neighbors show weaker results than other algorithms. Both in terms of execution time and the necessary resources, decision trees are superior to other algorithms. Also the main advantage of decision trees is their understandable detection of knowledge, that can be easily displayed to the user.

As a result of this research and the extracted knowledge, the operator will be able to accurately predict its customers' behavior, and will be able to direct their policies towards customers and their retention. At the same time, the results could lead to cost savings and better building of the company's budget.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje.

References

[1] L. Miguel APM. "Measuring the impact of data mining on churn management." *Internet Research*, vol. 11, no. 5, pp. 375–387, 2001.

[2] J. Hadden, et al. "Churn prediction: Does technology matter." *International Journal of Intelligent Technology*, vol. 1, no. 2, pp. 104–110, 2006.

[3] G. S. Linoff, and M. J.A. Berry. *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons, 2011.

[4] S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006.

[5] J. J. Rahul, and U. T. Pawar. "Churn prediction in telecommunication using data mining technology." *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 2, 2011.

[6] H. Jiawei, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[7] W. Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[8] K. Coussement, and D. Van den Poel. "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers." *Expert Systems with Applications*, vol. 36, no. 3, pp. 6127–6134, 2009.