Postmodern Openings
ISSN: 2068-0236 | e-ISSN: 2069-9387
Covered in: Web of Sciences (WOS); EBSCO; ERIH+; Google Scholar; Index Copernicus; Ideas RePeC; Econpapers; Socionet; CEEOL; Ulrich ProQuest; Cabell, Journalseek; Scipio; Philpapers; SHERPA/RoMEO repositories; KVK; WorldCat; CrossRef; CrossCheck

2019, Volume 10, Issue 4, pages: 44-53 | doi:10.18662/po/93

Churn Analysis in a Romanian Telecommunications Company

Andreea DUMITRACHE[1], **Monica Mihaela MAER MATEI**[2]

[1] PhD Student, Academy of Economic Studies, Bucharest, Romania, dumitrache.andreea03@gmail.com

[2] Conf. Univ. Dr., Academy of Economic Studies, Bucharest, Romania, matei.monicamihaela@gmail.com

Abstract: Telecommunications is one of the sectors where the customer base plays a significant role in maintaining stable revenues, so special attention is paid to prevent their migration to other providers. Over time, businesses in the telecommunications industry have faced multiple threats of financial loss from migrating customers who want to leave their telecom service provider in exchange for other offers from competing companies. An effective prediction model of this action can not only be viewed as an insurance policy, supporting stable revenue, but also provides suggestions for database management so that potential migrant customers can benefit from personalized offers and services, depending on their profile, thus preventing their loss. The aim of this paper is to predict customers who are going to defect in a Romanian mobile telecommunications company. The churn analysis is developed for post-paid customers. We used logistic regression to predict churn and a solution based on smoothed bootstrap technique to correct for the drawbacks of imbalanced classes. In our study this procedure did not significantly improve the performance of the logistic classifier measured by AUC (Area Under the Receiver Operating Characteristic curve). So even after balancing the sample we still obtain a really reduced value of the AUC, making it difficult to correctly predict churn phenomenon on the available data set.

Keywords: *Churn; class imbalance; customer; telecommunications.*

How to cite: Dumitrache, A., & Maer Matei, M.M. (2019). Churn Analysis in a Romanian Telecommunications Company. *Postmodern Openings, 10*(4), 44-53. doi:10.18662/po/93

1. Introduction

The cost of gaining new customers could be five times higher than retaining the existing ones (Keller and Kotler, 2016). This is the reason why marketing strategies should focus on customer retention. The efficiency of these strategies depends on the capacity of predicting churn phenomenon. Undoubtedly an accurate prediction model will lead to increased profitability of the company.

Telecommunications is one of the sectors where the customer base plays a significant role in maintaining stable revenues, so special attention is paid to prevent their migration to other providers. The customers switching from the current service provider to another are known as churners.

The present paper brings together common aspects of many areas of the market by focusing on studying the phenomenon of consumer migration, known in the specialty literature under the name of churn. Many of them are confronted with this fact because in most industries the market penetration of users (from mobile networks, for example) has exceeded 100%, becoming more and more dynamic and competitive. This problem has become central in many postmodern studies within the information and academic society. The acquisition of new users is practically impossible, as there are no new users, there are only users of rival companies that are exposed to numerous carefully designed marketing campaigns to win them (Shaaban et al., 2016).

The aim of this paper is to predict customers who are going to defect in a Romanian mobile telecommunications company. The churn analysis is developed for postpaid customers. An image of the scale of the churn phenomenon in Romania is provided by ANCOM report, which states that by the end of 2016, in Romania were 22.9 million active mobile services users, of which 11.5 million were subscription-based users. The telephone portability, a service made available by ANCOM, allows users to keep their phone number when changing service providers, thus increasing their freedom of choice and giving them the opportunity to enjoy the benefits of a competitive market. Since launching portability in 2008 and by the end of 2016, more than 3.5 million users have benefited from this service, of which 2.9 million are mobile users.

The number of studies published in the last years regarding churn prediction in telecommunications proves that this issue became a major concern. Up to now, the Romanian datasets were not the subject of the published articles in the postmodern academic world.

2. Literature review

Generally, the information used to build models for churn prediction in mobile telecoms industry includes customer demographics such as age, gender or tenure, contractual data, call-details, complaint data, billing information (Keller and Kotler, 2016; Huang et al., 2012).

At the end of the twentieth century, scientists focused on theoretical models for predicting the action of churn for improving data extraction methods and modernizing algorithms. In the 21st century, but also at the end of the 20th century, in postmodernism, the focus is on the scientific research on the data mining applications and their implementation in the real environment. Churn prediction has become a challenge, a critical business situation and one of the themes of the future. Predictive modeling based on pattern discovery is an intensely debated topic in the knowledge society because it presents multiple typical scientific innovations. This is an approach used to facilitate customer retention more efficiently and proactively (Berry & Linoff, 2004).

Data mining showed its applicability for predicting churn behavior in many countries. Decision trees and neural networks proved their utility for telecom churn management in Taiwan (Huang et al., 2012). Decision tree and logistic regression models were used to analyze churn for a data set provided by a UK mobile operator. In the model evaluation section was shown that classification trees outperform the logistic regression (Hassouna et al., 2016). Customer churn prediction for American telecom companies was conducted with artificial neural networks (Tsai et al., 2009). Empirical results showed that support vector machines (SVM) is very promising for churn prediction, performing better than Naive Bayes, Artificial Networks or decision trees (Zhao et al., 2005). A classifier based on Bayesian belief network identified factors affecting churn in telecommunication industry from Turkey (Kisioglu et al., 2011). A constant concern is related to the predictive performance improvement of these classifiers. One of the latest solutions builds a hybrid algorithm based on logistic regression and decision trees (De Caigny et al., 2018).

Given the scale of this problem, topical methods designed to improve the speed of the traditional ones were proposed in the past few years. For example, an approach that uses social network data was proposed and applied in Peru (Galvan et al., 2015).

In the context of churn prediction in telecommunications, customer churn is a rare event. In all data sets the number of churners is significantly outnumbered by the non-churners. All the classifiers parametric or

nonparametric are compromised by unbalanced data. Of course, logistic regression is no exception, its use on skewed data is not advisable because it will underestimate the probability of rare events. The implications of imbalanced data on logistic regression estimations are discussed in the literature (Hung et al., 2006). In a nutshell, classification results are overwhelmed by the dominant events and information coming from the rare cases is ignored. This issue was addressed by re-sampling the original dataset. One of the most popular sampling method is known as SMOTE: Synthetic Minority Over-Sampling Technique (Chawla et al., 2002). As suggested by its name, the minority class is over-sampled by creating "synthetic" examples. The new instances represent linear interpolations between original examples and randomly selected nearest neighbor's. A lot of other improved variants of SMOTE have emerged since its publication (Zhu et al., 2018). The efficiency of the sampling technique depends on the specificity of each domain. For example, the experiments undertaken in the context of churn prediction in telecom proved that when using AUC (Area Under the Receiver Operating Characteristic (ROC) curve) to evaluate the performance of classifiers, SMOTE does not significantly improve logistic regression results. Another finding of the paper is related to the selection of the sampling rate. The recommendation is to choose class ratio 2:3 (minority vs. majority) when AUC is taken into consideration (Zhu et al., 2018).

3. Methodology and data

In this article we analyze churn behavior on a sample of 10701 subscribers randomly selected from a database of a large telecoms company operating on the Romanian market. The migrating clients are marked with 1 in the Churn variable. We can classify the variables into four major categories:

• Demographic data, where we have information about the living area, age and gender.

• Details about the customer's lifecycle in the company: his seniority using the services of the company expressed in months (Duration), the number of months since changing last offer on the account (MonthsO), the number of months since the customer was called by telesales to change the subscription terms for another type of service (MonthsC).

• Data relating to the financial strength of each client, information expressed by the average of the three-month invoice that each customer has to pay, expressed in euro (Invoice), the value of the extra cost paid on services used outside the network (ExtraCosts).

- Information about subscriber's interaction with clients of competing networks: the number of national minutes made and received synthesized in two variables (MinC and MinR).

The data set we are analyzing depicts a class imbalance situation which means that the churners' class is rare over the sample. In this paper we are going to use a solution based on smoothed bootstrap technique referred to as ROSE (Random Over Sampling Examples). In order to balance class distribution new artificial data is drawn from the two conditional kernel density estimates of the classes. We have chosen this approach because it has shown excellent performance compared to other existing methods and also provides useful tools for model assessment phase (Menardi & Torelli, 2014).

As shown in the literature, the estimators of the performance measure should be reconsidered in imbalanced learning. The accuracy metric we are going to use is ROC curve and implicitly AUC. Due to the scarcity of data we prefer to train the model on the entire data set instead of splitting it in a train set and a test set. Using ROSE allows us to do that. The accuracy measures were obtained using an evaluation based on leave k out cross validation estimator. Therefore, the logistic regression is estimated on a ROSE sample from which at each round we exclude a small group of size 5. The prediction is made on the excluded set. These steps are repeated by a number of times equal to sample size over 5 (Lunardon et al., 2014).

We have used a partitioning based cluster method in order to analyze the churn phenomenon on more homogeneous sets. We have built 3 groups of clients using k-means clustering on two features: the average of the three-month invoice and the value of the extra cost paid on what it uses outside the network. The distributions of these two variables show a very high variance compared to the other customer characteristics.

4. Results

Mainly the heterogeneity of the sample comes from the two features mentioned above: Invoice and extra cost. The following two figures emphasize this issue.

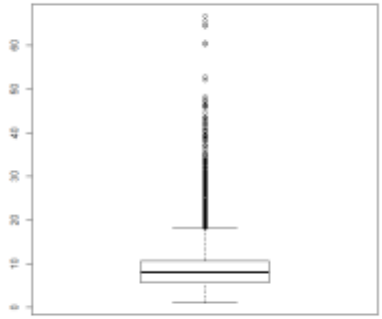

Figure 1. Distribution of average invoice amount

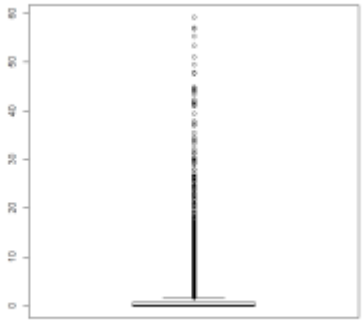

Figure 2. Distribution of extra cost

Both variables present a right skewed distribution that led to splitting the data set into more homogeneous customer groups. The churn behavior was analyzed inside these groups emerging from a k-means method. The solution has an average silhouette width of 0.54. As shown in the next chart, the variable measuring the average of the three-month invoice has a higher discriminative power. Cluster 1 takes over 67% of the observations in our sample and it depicts a high level of homogeneity compared to the other two clusters.

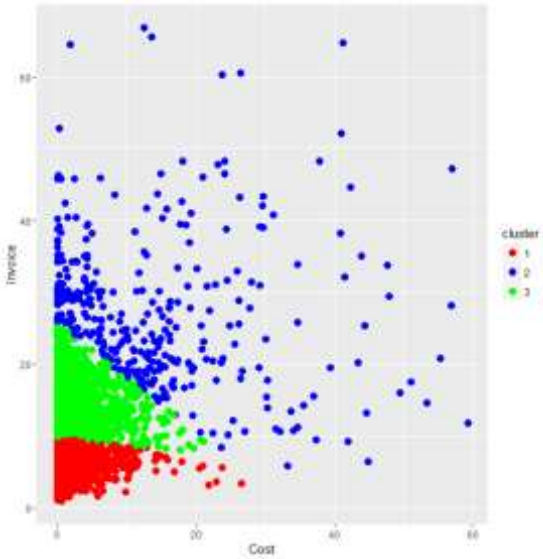

Figure 3. Customers Clusters

The objects belonging to Cluster 3 represent 29% of the initial sample. In these two clusters we do not find customers whose features (Cost and Invoice) take values in the upper limit of their distributions. Cluster 2 shows a really high variance, most of the objects classified here having an outlier behavior. As a consequence, we did not find a statistically significant model for this segment. The churn rate for the first group (represented with red) is 14.8%, for the second one (blue) is 24% and in the third cluster (green) the churn rate is 17%. Therefore, the churn rate increases as the Invoice value and the cost go up.

Table 1. Logistic regression results for Cluster 1

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)	Exp(coef)
(Intercept)	-0.763	0.347	-2.198	0.027	0.466
Duration	-0.009	0.002	-4.847	1.25E-06	0.99
MonthsO	0.012	0.001	10.753	<2e-16	1.012
MinR	0.537	0.162	3.319	0.0009	1.710

Because we wanted to test the efficiency of ROSE on our data set, initially we did the estimation using the traditional pattern. This means we have randomly selected 70% of the customers from cluster 1 and used them to train the classifier. The other subset (30%) was used to test the accuracy of the model. The results are presented in Table 1 and figure 4. The features influencing the probability of a customer to churn are: the length of the period the customer used the company services, the number of months since he changed his last offer, the number of calling minutes outside the network.

The former variable has the highest impact: for every 1-minute increase in the number of calling minutes the odds of churning is increased by 71%. For the other variables we did not find significant differences between the distribution of churners and non-churners. The AUC value is very low (0.604) showing low discrimination power.

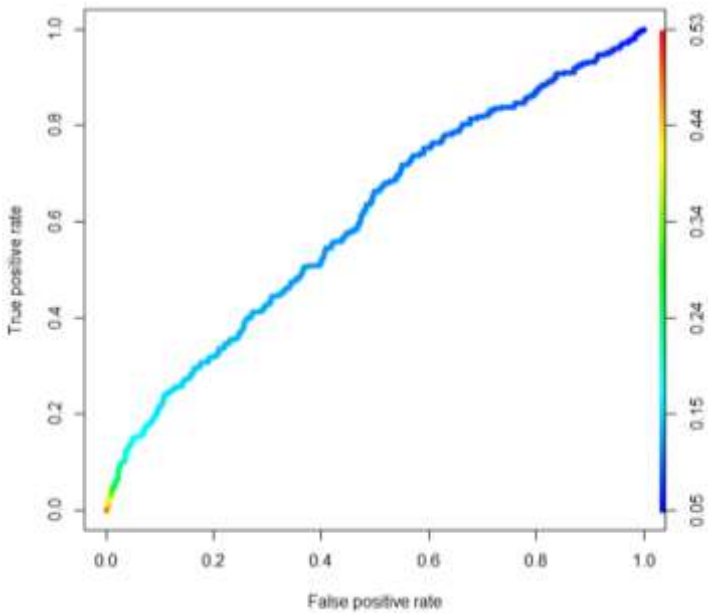

Figure 4. ROC curve for cluster 1

Also, as seen in figure 4, given the class imbalance problem, the threshold used for classification is situated well below the default value of 0.5.

Using ROSE with a class ratio of 2:3 the leave K out cross-validation estimate of AUC is 0.625. So even after balancing the sample we still obtain a really reduced value of the AUC, making it difficult to correctly predict churn phenomenon. Our findings are consistent with other experiments undertaken in the telecom churn context showing that resampling does not improve the performance of the logistic classifier.

We have followed the same procedure for cluster 3. The results are shown in Table 2. The value of the extra cost paid on services used outside the network showed significance for this customer segment.

Table 2. Logistic regression results for Cluster 2

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)	Exp(coef)
(Intercept)	-1.137	0.497	-2.286	0.022	0.321
MonthsO	0.031	0.004	6.455	1.08E-10	1.031
Duration	-0.007	0.002	-2.597	0.009	0.992
Cost	0.036	0.021	1.690	0.091	1.037

The AUC value for this model is 0.64 and after using the resampling method slightly increased to 0.668.

5. Conclusions

In order to overcome the shortcomings encountered when training classifiers on unbalanced samples we used a solution based on smoothed bootstrap technique referred to as ROSE (Random Over Sampling Examples). As proved before in other postmodern studies, in the telecom churn context resampling does not improve the performance of the logistic classifier. So even after balancing the sample we still obtain a really reduced value of the AUC, making it difficult to correctly predict churn phenomenon. In further study other classifiers will be tested and the features set will be expanded with historical data.

References

Berry, M. J. A., Linoff, G. S. (2004). *Data mining techniques second edition – for marketing, sales, and customer relationship management*. 2nd ed., ISBN 0-471-47064-3, United States of America

Chawla NV, Bowyer KW, Hall KO and Kegelmeyer WP (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(3): 321–357.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research. European Journal of Operational Research, 269* (2), 760-772, https://doi.org/10.1016/j.ejor.2018.02.009.

Shaaban, E., Hassanien, A. E., (2016). Churn Prediction Retention Framework. *International Journal of Advances in Computer Science & Its Applications*, (6)1, 11-16, ISSN 2250-3765.

Galvan, A. R. M., & Navarro, K. R. C. (2015). Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies. In *Information Management and Big Data* (pp. 120-132). Springer, Cham.

Hassouna, M., Tarhini, A., Elyas, T., & AbouTrab, M. S. (2016). *Customer Churn in Mobile Markets a Comparison of Techniques*. arXiv preprint arXiv:1607.07792.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524, doi10.1016/j.eswa.2005.09.080.

Keller, K. L., & Kotler, P. (2016). *Marketing management*. Pearson.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert Systems with Applications*, *38*(6), 7151-7157

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, 6(1),79-89 DOI: 10.32614/RJ-2014-008

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92-122.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547-12553.

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. In *International Conference on Advanced Data Mining and Applications* (pp. 300-306). Springer, Berlin, Heidelberg.

Zhu, B., Baesens, B., Backiel, A. E., & vanden Broucke, S. K. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, *69*(1), 49-65.