The Eleventh International Conference on Digital Information Management (ICDIM 2016)

Churn Perdiction in the Telecom Business

Georgina Esteves
Department of Informatics Engineering,
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
Email: ei10010@fe.up.pt

João Mendes-Moreira
LIAAD-INESC TEC
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n 4200-465, Porto, Portugal
Email: jmoreira@fe.up.pt
Telephone: (351) 225082142

***Abstract*—Telecommunication companies are acknowledging the existing connection between customer satisfaction and company revenues. Customer churn in telecom refers to a customer that ceases his relationship with a company. Churn prediction in telecom has recently gained substantial interest of stakeholders, who noticed that retaining a customer is substantially cheaper that gaining a new one. This research compares six approaches using different algorithms that identify the clients who are closer to abandon their telecom provider. Those algorithms are: KNN, Naive Bayes, C4.5, Random Forest, AdaBoost and ANN. The use of real data provided by WeDo technologies extended the refinement time necessary, but ensured that the developed algorithm and model can be applied to real world situations. The models are evaluated according to three criteria: are under curve, sensitivity and specificity, with special weight to the first two criteria. The Random Forest algorithm proved to be the most adequate in all the test cases.**

I. Introduction

Since the 1990s the telecommunications sector became one of the key areas to the development of industrialized nations. The main boosters were technical progress, the increasing number of operators and the arrival of competition. This importance has been accompanied by an increase in published studies and by new marketing strategies [1].

In order to acquire new costumers, a company must invest significant resources to provide a product or service that stands out from the competitors. However, continuously evolving the product itself is not enough. Companies are realizing that a customer-oriented business strategy is important for sustaining their profit and preserving their competitive edge [4]. As acquiring a new costumer can add up to several times the cost of efforts that might enable the firms to retain a customer, a best core marketing strategy has been followed by most in the telecom market: retain existing customers, avoiding customer churn [2], [3].

Several studies have been done in order to test one or two methods for churn prediction. There is also a study [11], that compares a large set of algorithms. This study differs from the one we present by using several relatively small datasets including also demographic data. However, demographic data is not always accessible to third parties software providers, such as the ones that present solutions for churn prediction. The data that will be used during this work assures the quality of the final solution. The dataset is an extended collection of calls conducted by real customers, ensuring that the developed algorithm can be applied to real world big datasets even without access to demographic data of the clients.

A revision of the works on churn prediction in telecom companies is done in Section II. Then the dataset (Section III) and the experiments (Section IV) are described followed by the conclusions and a discussion on future work (Section VI).

II. Churn prediction in the telecom industry

According to the author [5], 'costumer churn' in telecom business refers to the costumer movement from one provider to another. 'Customer management' is the process conducted by a telecom company to retain profitable costumers. The continuous evolution of technology has opened up the telecommunications industry, making this market more competitive than ever. These companies are realizing that a customer-oriented business strategy is important for sustaining their profit and preserving their competitive edge [4]. As acquiring a new costumer can add up to several times the cost of efforts that might enable the firms to retain a customer, a best core marketing strategy has been followed by most in the telecom market: retain existing customers, avoiding customer churn [3], [2].

Two main types of targeted approaches to manage customer churn [4] were identified: reactive and proactive. In the reactive approach, the company waits until the customer asks to cease their contract to act. In this situation, the company will then offer some advantages and incentives to retain the customer. The other approach is the proactive approach, where a company tries to identify which customers are more likely to churn before they do so. In this case, the company provides special offers to keep them from churning.

A few studies have been conducted to evaluate which approach, reactive or proactive, is better to a telecom company. The majority of them agreed that the proactive approach achieves better results [12], [13], and the work presented in this paper is based in this second approach. The main objective is to predict beforehand the customers more likely to churn using data mining techniques.

A distinction between types of churn must be made. Involuntary churn is when circumstances outside the user and service providers control affect the decision of ceasing their relationship. Customers relocation to a distant location and death are part of this kind of churn. Voluntary churn is when

978-1-5090-2641-8/16/$31.00 ©2016 IEEE

a customer actively decides to leave the product or service of a company. Involuntary churn tends to be discarded to churn prediction, because those are the ones that do not represent the company-customer relationship.

A. Churn prediction approaches

The ability to predict that a particular customer has a high risk of churning, while there is still time to do something about it, represents an opportunity of additional revenue source for every business. So, investment has been conducted in this area, and multiple approaches have been studied and tested.

In the telecom business, the majority of the studies done on churn prediction focus on applying only one or two data mining methods.

One of the studied approaches [16] consisted on applying SVM to churn prediction in subscription services. The objective was to construct an accurate churn model using and tuning this technique. The customer churn prediction performance of the model was benchmarked to logistic regression and random forest. The authors chose as the main metric to evaluate their models Area under curve (AUC). The best value regarding their SVM models was 85.14, and the random forest model used as benchmark conquered a final AUC value of 87.21.

This approach [15] evaluates the efficiency and the performance of Decision tree and Logistic regression techniques. They concluded that the accuracy achieved with decision tree was higher than the logistic regression technique.

Another of the investigated studies [17] consists on comparing three classification techniques: Logistic regression, automatic relevance determination (ARD) Neural Networks and Random Forests. The best fit in this study was the Random Forest model, with an AUC value of 83.19. The ARD NN model reached a similar value, with an AUC of 83.10.

Only a few authors [11] focused on comparing multiple strategies to predict customer churn. The authors chose 21 algorithms to model, and were evaluated regarding their AUC and percentage of correctly classified (PCC) instances, in both training and testing sets. Those algorithms include SVM with different kernels, C4.5, Random Forest and KNN. The highest AUC value was obtained by the Random Forest model in both train and test sets, with values 0.8249 and 0.8319 respectively.

III. Dataset

The models that predict customer churn are based on knowledge regarding the company's clients and their calls. That information is stored in a database table and is called dataset.

All the models were trained and tested with this data. But before that can happen, this collection had to go through multiple transformations and pre-processing techniques to make it suitable to predict upon.

A. Data collection

For the purpose of this research, real data from WeDo technologies client calls was provided by the company. The data is stored in a SQL file, and has a size of more than 131 gigabytes. The file contains one table with over 1.2 billion entries from 5 million different clients.

B. Data selection

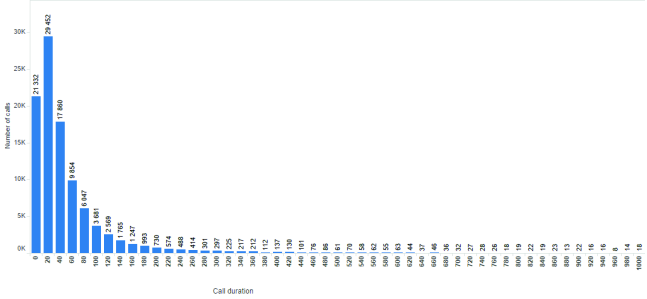
Due to computational and time limitations, the dataset available to this research was too large and needed to be sampled. In order to prevent the lost of valuable information and keep the final results accurate, a simple selection of the first N table entries was not viable. The right approach to this problem is to retrieve all the call records from a group of clients. The final dataset contained over 100 thousand calls from 160 clients dated between 30 June 2012 and 31 January 2013.

C. Data analysis

The dataset available to this research contains information regarding call information of telecom company clients.

Table I contains an overview of the data. It identifies the variables, their values and a simple description of their meaning. Figure 1 is a visual representation of the duration in minutes of client calls. The calls were grouped into 20 minutes intervals. The graph was also trimmed at 1000 minutes to make the graph more easily understandable. We can see that the majority of the calls on our dataset has a duration between 1 and 60 minutes. This number continues decreasing with the growth of the call duration, and stabilizes near 440 minutes.

Fig. 1. Number of calls per call duration.



In our dataset, variable "direction" expresses if a call record regards an incoming call ("I") or an outgoing call ("O"). The majority of the calls represent an outgoing call, and only around 35 thousand are incoming calls.

Around 90% of the calls belong to the Mobile and On-Net groups.

The majority of the calls are between numbers from the same country, named local (L) calls. Only 3% of the calls in our dataset are international (I). A similar result can be found when analyzing dropped calls data. A call is said to be dropped when due to technical reasons, is cut off before the speaking parties had finished their conversation and before one of them had hung up. Only around 2% of our data has this specific characteristic.

But the main problem regarding our data was the lack of labels informing whether the costumers churned and the ones that did not. According to some authors [6], [7], we can state that a costumer has churned a telecom company when he does

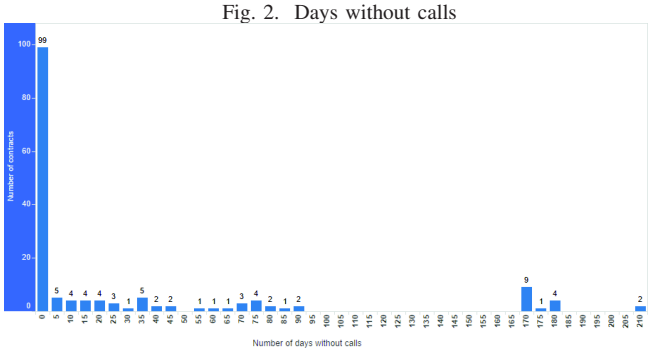
TABLE I
VARIABLES IN THE DATASET

Variable	Value	Description
DATE	YearMonthDay	Date of the call.
TIME	HoursMinutesSeconds	Time of the call.
DURATION	Seconds	Call duration.
MSISDN	Numeric	Anonymized number. If Incoming, it is the number getting the call. If Outgoing, is the number calling.
OTHER_MSISDN	Numeric	The ”other” number in the call (regarding the previous variable).
CONTRACT_ID	Numeric	Client code.
OTHER_CONTRACT_ID	Numeric	Not present in all data (does not belong to the operator).
START_CELL_ID	Numeric	Should represent the calling device (Outgoing).
END_CELL_ID	Numeric	Should represent the device getting the call (Ingoing).
DIRECTION	I, O	Represent a incoming call,(”I”) or outgoing,(”O”).
CALL_TYPE	FI, MO, ON, OT, SV	FI - The other device in the call belongs to a wireline. MO - The other device in the call belongs to a mobile network. ON - On-Net, both incoming and outgoing systems belong to this operator. OT - Others. SV - Services:VoiceMail calls, etc.
DESTINATION_TYPE	I, L	Local (L) or international (I).
DROPPED_CALL	Y, N	Dropped call (Yes or No).
VOICMAIL	Y, N	Call went to voicemail (Yes or No).

not do or receive any communication during 30 days or more. We used this criterion to label the instances.

The respective calculations were conducted, and final results are shown in figure 2. More than 60% of clients have done communications in less then 5 days, which tells they are currently active. Regarding churn itself, 26% of the customers in the data churned the company. We also have an estimate of approximately 26% customers that churned the company, with a maximum day difference of 210 days without communications.

Fig. 2. Days without calls



The final distribution of the clients concerning churn had a 16% churn to 84% not churn ratio. This is our target variable, which means that the objective of the developed models is to predict the outcome of this variable through the study of the other variables in our data. The problem regarding imbalanced class will be addressed in the next chapter.

IV. EXPERIMENTAL SETUP

To guarantee the results integrity when comparing multiple algorithms, we need to assure that all the algorithms are tested in the same conditions and with the same data. Variables ”contract_id” and ”msisdn” had to be removed from the data that was going to train the multiple models, because they have a direct connection with the outcome of target variable (identify the client itself).

The solution found consists in dividing the original data into two subsets: train set and test set with a 70/30 ratio.

As we have already mentioned on the previous section, our target class is imbalanced, and this value disparity could have a significant negative impact on the final models regarding model fitting. The chosen approach was a hybrid sampling method [8] named Synthetic Minority Over-sampling Technique (SMOTE) that was applied to the training set. This method approaches the data in two ways: generates new examples of the minority class using the nearest neighbors of that cases, and under-samples the majority class.

The evaluation metrics chosen are Area Under Curve (AUC), sensitivity and specificity. In customer churn analysis it might be more expensive to incorrectly infer that customer is not churning then to give a general reduction in prices for services to clients that are not planning to leave the company. Since the priority in our case study is given to identifying churn clients rather than not churn ones, Sensitivity is more relevant than Specificity in our results.

The algorithms applied were chosen due to their diversity of representation and learning style, and their common application on this kind of problems. We also took into consideration studies regarding the popularity and efficiency, as well as some studies that compare data mining approaches in classification problems [9], [14].

Six different machine learning models were trained and compared among themselves. Those algorithms were:

- K-Nearest Neighbor (Knn);
- Naive Bayes;
- Random Forest;
- The C4.5 decision tree method;
- AdaBoost;
- Artificial Neural Networks (ANN);

Each model was tuned and evaluated using 3 repeats of 10-fold cross validation, a common configuration on data mining for comparing different models [10]. The model with the best scores is then chosen to make predictions in new data, defined as test set.

A random number seed is defined before the train of each one of the algorithms to ensure that they all get the same data partitions and repeats.

In order to acquire the most adequate model to our problem, there is a need to explore which are the best parameters for each algorithm. To do so, we constructed graphs with the performances of different algorithm parameter combinations, with the final objective of finding trends and the sensitivity of the models.

In the KNN algorithm, We studied the effect of parameter *k* variation in the ROC and Sensitivity results, with the final objective of selecting the one who reached the highest scores on those metrics.

Regarding the Naive Bayes algorithm a study was conducted around the influence of the kernel choice in the models behavior. When the "use kernel" value is set as TRUE, a kernel density estimate is used for density estimation; if it is set as FALSE a normal density is estimated.

In the algorithm Random Forest, tuning parameter "mtry" represents the number of variables randomly sampled as candidates at each split of the decision tree. Multiple values of this parameter were tested to discover what value of the mtry parameter returns the most adequate model to our case study. 500 trees were constructed to train the model.

On the Adaboost algorithm, a study regarding which value combination of two tuning parameters returned the best model according to our evaluation. Those tuning parameters were number of trees to be generated (iter) and maximum tree depth (maxdepth).

In the ANN predictive model two parameters were tuned to find the best fit to our problem: size and decay. The size value represents the number of units in the hidden layer, and the decay value is a parameter used in weight decay formula that penalizes solutions with high weights and bias.

All the computation during this research was done with the resources at FEUP Grid, and using a library for parallel computation.

V. Results

In order to acquire the most adequate model to our problem, there is a need to explore which are the best parameters for each algorithm. To do so, we constructed graphs with the performances of different algorithm parameter combinations, with the final objective of finding trends and the sensitivity of the models.

The models were trained using 89159 entries, 10 predictors and 2 classes.

The results are evaluated under the following hypothesis:

- **H0:** all algorithms have similar performances.
- **H1:** there are differences in the performances of the different algorithms.

In order to compare multiple algorithms, the general recommended methodology is as follows. First, to apply an omnibus test to detect if at least one of the algorithms performs different than the others. Second, if we find this difference difference, apply a pair-wise test with the corresponding post-hoc correction for multiple comparisons. We chose Friedman rank test [18] as our omnibus test, and Nemenyi [19] to our post-hoc test.

The results of the Friedman rank test were the following.

Friedman rank sum test

Friedman chi-squared $= 4930.2, \text{df} = 5, \text{p-value} < 2.2e-16$

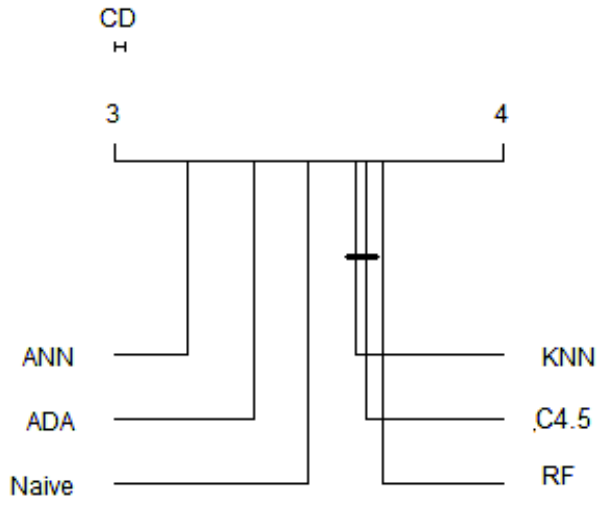
The p-value shown above denotes that there is at least one algorithm that performs differently than the rest. Therefore, we can proceed with the post-hoc analysis of the results with the Nemenyi test. Alpha was set at 0.05.

Nemenyi test

Critical difference $= 0.02525, \text{k} = 6, \text{df} = 534950$

The critical difference value is 0.02525, which means that the difference between two algorithms must be higher than that value to guarantee that we can refute the null hypothesis.

Fig. 3. Critical Difference Diagram



By the analysis of graph 3 we can state that we cannot exclude the hypothesis that the average ranks of KNN and Random Forest algorithms are equal. The differences among the remaining algorithms are higher than the critical difference value,

The AUC is a common evaluation metric for binary classification problems. It represents the area value created by the ROC curve. If a classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is similar to random guessing, the true

TABLE II
AUC VALUE COMPARISON

	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
KNN	0.951	0.953	0.954	0.955	0.957	0.960	0
NB	0.687	0.699	0.705	0.704	0.708	0.717	0
C4.5	0.928	0.933	0.935	0.935	0.937	0.945	0
RF	0.988	0.990	0.990	0.990	0.991	0.992	0
ADA	0.768	0.777	0.781	0.781	0.786	0.797	0
ANN	0.690	0.708	0.713	0.713	0.719	0.732	0

positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5. The values regarding AUC for our trained model are on table II. Among the studied models, three of them achieved very good AUC values: KNN, C4.5 and and Random Forest.

TABLE III
SENSITIVITY VALUE COMPARISON

	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
KNN	0.871	0.879	0.882	0.882	0.885	0.893	0
NB	0.548	0.565	0.569	0.569	0.575	0.586	0
C4.5	0.845	0.861	0.865	0.864	0.869	0.875	0
RF	0.897	0.902	0.905	0.905	0.908	0.911	0
ADA	0.404	0.414	0.435	0.433	0.446	0.472	0
ANN	0.348	0.386	0.437	0.427	0.458	0.485	0

Regarding Sensitivity values displayed on table III, the results were worse than in the previous metric. The same models that had the higher AUC values are still on the top Regarding Sensitivity values. The AdaBoost and ANN models have very poor Sensitivity values, and can be considered not a good fit to our problem.

A. Variable importance

We conducted a study to acquire these values for the Random Forest method, to assess which predictors had the largest impact on the model which got the best results.

The results of this study are presented on table IV. All measures of importance were scaled to have a maximum value of 100.

Variable "duration" has clearly the most importance to our model, setting the top value of 100. The next variables significant to our model are the call "direction" and a number representing the other person on the call.

TABLE IV
ROC CURVE VARIABLE IMPORTANCE

Variable	Importance
duration	100.000
direction	61.275
other_msisdin	55.965
start_cell_id	29.924
dropped_call	24.278
destination_type	24.205
voicemail	22.547
end_cell_id	12.009
call_type	4.723
other_contract_id	0.000

VI. CONCLUSION

The importance of this type of research to the telecom market is continuously growing. Data collection is becoming an everyday task to all companies, and the value of that data can come from multiple sources. Churn prediction is becoming one of those sources that create revenue to the company. Being able to prevent when clients are going to cease their contract with the company opens the possibility of renegotiating that contract in order to retain the costumer.

Although multiple studies have been conducted in this area, most of them use only one or two algorithms. In this way, there is a gap in knowledge regarding the comparison on efficiency of several algorithms to the same problem and upon the same experimental setup using a large dataset. This research compares multiple approaches to predict customer churn, identifying which ones are the most fitted to the original problem.

This research aimed to create suitable models that predict customer churn. These models needed to register high values in the defined metrics: AUC and Sensitivity. To validate the models, we chose to implement 10-folds cross validation with 3 repeats.

The data itself was not suitable to predict upon. It had to go through multiple transformations to make it fit to train and test the models.

As the database was provided by a company with their real values, it did not contained the information regarding the target variable: if a costumer had churned or not. According to the standard in this circumstances, we set that variable according to a simple rule: if an account did not make or receive any calls during a 30 day time period, we considered that client has cease the contract. However, when analyzing this new variable we came across another problem: it was imbalanced. Only around 18% of the entries represented churning clients. If our models were trained upon this data, it could result in models that disregard the minority class. Our approach to solve this problem consisted on applying an hybrid sampling method SMOTE. Our final dataset was much more balanced than before with the churned class represented by 0.43% of all entries.

Six models were trained and tested. The algorithms themselves were chosen due to their diversity and applicability in this kind of prediction. Those algorithms were KNN, naive bayes, random forest, C4.5, AdaBoost and ANN.

Random forest model achieved the best results in both ROC and Sensitivity measurements. With a highest ROC value of 0.9915 and Sensitivity value 0.9110, it distinguishes itself from the other models.

KNN and C4.5 models come in second and third place respectively. They acquired similar results to the random forest model, being their AUC and Sensitivity values slightly lower.

The remaining algorithms, (naive bayes, adaboost and ann) failed to positively predict upon our data. Their ROC values were acceptable, but their Sensitivity values are considerably low. This means that these algorithms did not conceived good models according to our metrics and needs.

As the dataset used in this research represent real calls from a telecom company, this model can effectively be used to the company's welfare. The administration, using this kind of model, can predict and target the clients that are close of ceasing their contract.

Although we consider the conducted study a success, we acknowledge that there is still room for improvements.

- Since our dataset timeframe is 7 months (difference between first and last entry), some of the older data can influence in a negative way the models. It would be interesting to investigate which data timeframe produce the best result and how the efficiency of the models is influenced by this timeframe.
- All the algorithms suffered parameter tuning to identify which parameter values returned the best models. But there are many other approaches to tune the models that were not tested and could improve the final models.

ACKNOWLEDGMENT

This work is financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT Fundao para a Ciłncia e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

The authors would like also to thank WeDo Technologies for the data provided for this study.

REFERENCES

[1] Gerpott, T. J., Rams, W., and Schindler, A.: Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy, 25(4):249269. (2001)

[2] Kim, H. S. and Yoon, C. H.: Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, 28(9- 12):751765. (2004)

[3] Kim, M. K., Park, M. C., and Jeong, D. H.: The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy, 28(2):145159. (2004)

[4] Tsai, C. F. and Lu, Y. H.: Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10):1254712553. (2009)

[5] Yen, C. and Wang, H.-y.: Applying data mining to telecom churn. 31:515524. (2006)

[6] Oentaryo, R. J., Lim, E.-p., Lo, D., Zhu, F., and Prasetyo, P. K.: Collective Churn Prediction in Social Network. (2012)

[7] Radosavljevik, D., Putten, P. V. D., and Larsen, K. K.: The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications : What to Predict , for Whom and Does the Customer Experience Matter 3(2):8099. (2010)

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE : Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16:321357. (2002)

[9] Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., and Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems. (2008)

[10] Kuhn, M.: Building predictive models in r using the caret package. Journal of Statistical Software. (2008)

[11] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B.: New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1):211229. (2012)

[12] Retana, G., Forman, C., and Wu, D.: Proactive customer education, customer retention, and demand for technology support: Evidence from a field experiment. Manufacturing and Service Operations Management, 18(1):3450. (2016)

[13] Olaleke, O., Borishade, T., Adeniyi, S., and Omolade, O.: Empirical analysis of marketing mix strategy and student loyalty in education marketing. Mediterranean Journal of Social Sciences, 5(23):616625. (2014)

[14] Fern, M. and Cernadas, E.: Do we Need Hundreds of Classifiers to Solve Real World Classification Problems ? Journal of Machine Learning Research, 15:31333181. (2014)

[15] Dahiya, K., and Bhatia, S.: Customer Churn Analysis in Telecom Industry. (2015)

[16] Coussement, K., and Poel, D. Van Den: Churn prediction in subscription services : An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34, 313327. (2008)

[17] Buckinx, W., and Poel, D. Van Den: Customer base analysis : partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164, 252268. (2005)

[18] Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics 11, 8692. (1940)

[19] Nemenyi, P.: Distribution-free multiple comparisons. Ph.D. thesis, Princeton University. (1963)