Transactions on Machine Learning
and Data Mining
Vol. 3, No. 2 (2010) 80-99
© ISSN: 1865-6781 (Journal),
ISBN: 978-3-940501-19-6,
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter?

Dejan Radosavljevik[1], Peter van der Putten[1], Kim Kyllesbech Larsen[2]
[1] Leiden Institute of Advanced Computer Science,
Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
[2] International Network Economics, Deutsche Telecom AG,
Landgrabenweg 151, D-53227 Bonn, Germany
{dradosav, putten}@liacs.nl
kim.larsen@t-mobile.nl

Abstract. Prepaid customers in mobile telecommunications are not bound by a contract and can therefore change operators ('churn') at their convenience and without notification. This makes the task of predicting prepaid churn both challenging and financially rewarding. This paper presents an explorative, real world study of prepaid churn modeling by varying the experimental setup on three dimensions: data, outcome definition and population sample. Firstly, we add Customer Experience Management (CEM) data to data traditionally used for prepaid churn prediction. Secondly, we vary the outcome definition of prepaid churn. Thirdly, we alter the sample of the customers included, based on their inactivity period at the time of recording. While adding CEM parameters did not influence the predictability of churn, one variation on the sample and especially a particular change in the outcome definition had a substantial influence.

Keywords: Prepaid Churn, Data Mining, Customer Experience Management.

1 Introduction

The problem of churn, or loss of a client to a competitor, is a problem facing almost every company in any given industry. This phenomenon is a major source of financial loss, because it is generally much more expensive to attract new customers than it is to retain and sell to existing ones. Therefore, churn is important to manage, especially in industries characterized by strong competition and saturated markets, such as the mobile telecom industry. Prepaid subscribers of mobile telecom services are not bound by a contract; therefore, they can churn at their convenience and without notification, which makes the task of predicting the likelihood and moment of churn very important.

The first step in managing churn is identifying the customers with high propensity to churn. Published papers on churn modeling on real world datasets in mobile telecommunications are relatively scarce, due to the commercially sensitive nature of the problem. This is even more evident for papers based on European mobile telecom operators' data. Most of the available research relates to postpaid churn [1,2,3,4,5,6,7,8,9,10,11]. Even fewer studies are available for prepaid churn in this industry [12,13,14]. The majority of these assume a fixed, single experimental setup in terms of outcome definition, population and data parameters (see Section 2 for a more detailed overview of prepaid churn modeling literature). Their focus is mostly on the data mining algorithm used or algorithmic tuning. In order to understand prepaid churn modeling better, we decided to take an end-to-end view and test different experimental setups by varying on three dimensions: data, population sample and outcome definition [15]. We used standard data mining algorithms, decision trees (in their standard form) and logistic regression [16,17], because it was not our objective to determine the impact of the algorithm; research on data mining algorithms is abundant. Furthermore, from a business perspective, the output of either these algorithms (the model) is very easy to interpret and communicate to parties that do not have extensive experience with data mining (e.g. business managers). This quality is even more evident in the case of decision trees, which have a very intuitive graphical representation. Finally, as our experiments have shown, the choice of algorithm is a minor factor of influence compared to the other dimensions mentioned.

Making new types of data available for modeling may improve model performance. We provide an overall framework for measuring customer experience, and in the experiments we investigate the added value of customer level service quality metrics (dropped calls, call setup duration, SMS delivery failure rate etc).

The definition of a population sample and outcome is primarily driven by the business objectives, i.e. how the model will be used. In general, operators discontinue the service and consider prepaid customers churned from an administrative perspective if they have not had an activity (e.g. made a call, sent an SMS, etc.) for a period of six months. However, from a churn prediction and marketing perspective such a long period is not very useful. Usually, the customer has churned long before that time. Therefore, it is much better to use a shorter period of inactivity as an outcome definition of churn. Our variations in population sample are based on the customer's inactivity period at the moment of recording. The dilemma here is, if customers have been inactive already for one month at the time of recording, are they retainable, or have they simply thrown away the SIM card? What should be the maximum period of inactivity allowed at the time of recording?

The main contributions of this paper are the following. First, we provide a novel theoretical business framework for measuring customer experience in mobile telecommunications. Second, to our knowledge we are the first to investigate the added value of service quality oriented customer experience data for predicting prepaid churn. Third, we explore the impact of changing the criterion for the population sample. Fourth, we propose different outcome definitions of prepaid churn and measure the impact of the change in outcome definition on the model performance. Finally, this research was conducted by one of the leading mobile telecom operators in Europe, driven by high priority business needs and using large amounts of customer data, which gives it a real world dimension. Given the highly competitive, confidential and strategic nature of mobile telecommunications churn management, real world results (based on the European telecom market) are not often available in research literature.

The remainder of this paper is structured as follows: In Section 2 we present how (prepaid) churn modeling has been performed in the past. Section 3 provides an overview of Customer Experience Management. In Section 4, the research setup and the experimental design are discussed, followed by results in Section 5. We end the paper with a discussion (Section 6) and a conclusion (Section 7).

2 Prepaid Churn Modeling

In this section, we will present how modeling churn in mobile telecommunications has been addressed in previous research.

As mentioned in the introduction, the majority of papers on wireless churn assume a single experimental setup and deal with postpaid churn [1,2,3,4,5,6,7,8,9,10,11]. Some of these papers do not explicitly state that they are related to postpaid churn; nonetheless, this conclusion can be made from the datasets they use, which contain demographic information and contract data. These parameters are unavailable for prepaid subscribers, as there is no contract in the real sense of the word. This is the key difference between prepaid and postpaid churn prediction, even though quite a few similarities exist.

Prepaid churn modeling is typically done by performing analysis on aggregated Call Detail Records (CDRs), but additional factors, such as subscription details (e.g. duration, subscription or discontinuation of services) and handset data, are often included. Parameters used in previous papers are as follows.

Archaux et al. [12] take into account the following data: invoicing data, such as the amounts refilled by clients or amounts withdrawn by companies for the subscribed services and options; data relating to usage, such as the total number of calls, the share of local, national or international calls, consumption peaks and averages; data relating to the subscription, such as date of beginning (age of the subscription), the current tariff plan and the number of different plans the client used; data relating to application and cancellation of services; data related to the current and the previous profitability of the clients. Alberts [13] selects data only related to usage and billing information: average duration of a single incoming and outgoing call, ratio between outgoing and incoming call durations, sum of incoming and outgoing revenues, the current and the maximum number of successive non-usage months, cumulative number of recharges, average duration of incoming and outgoing calls over all past months, the number of months since the last voicemail call, the number of months since the last recharge, the last recharge amount, the average recharge amount of all past months, etc.

The algorithm is the focus of many of the data mining efforts related to churn. Papers related to postpaid churn have investigated the performance of standard data mining algorithms, such as decision trees [1,2,4,5,6,7,8,9,10,18], logistic regression [2,5,7,9, 10,18], neural networks [1,2,4,5,6,7,11], Bayesian classifiers [5] and support vector machines [18], and compared them to novel approaches.

In prepaid churn modeling, Archaux [12] compares the performance of models built using neural networks and support vector machines, while Alberts [13] compares models built using survival analysis with models built using decision trees. To summarize, in research literature there is a lot of emphasis on algorithm testing and tuning, whereas in real world practice this is a smaller piece of the puzzle with relatively modest impact. Hence, we decided to focus more on experimental setup.

Recently, the influence of Social Networks on (prepaid) churn has entered the focus of interest of both business and academic communities [14,19,20]. These papers strive to answer the question whether the decision of a subscriber to churn is dependent on the existing members of the community with whom the subscriber has a relationship. Dasgupta et al. [14] show that diffusion models built on call graphs (used for identification of Social Networks) have superior performance to their baseline model. We consider this approach to be a very progressive novelty, with relation to both algorithm and data. However, it is our opinion that these results are biased by the fact that their baseline model is constructed on CDR data alone, without including other important factors, such as the handset, prepaid account balance or inactivity period. In our opinion, if these variables would have been added in their baseline model, the improvements of their approach would not have been as high.

As we will show in our experiments, combining past behavior (extent of usage) with current factors (e.g. prepaid account balance, phone type, price plan etc) is crucial for predicting future behavior (churn). In general, there is a gap between traditional methods that do not take the social networks into account and social networks techniques that only mine the network. In our future work we are planning to investigate the value of more hybrid approaches, but this is out of scope for this article.

In addition, we consider the origin of the data used in previous research. As telecom markets differ in various parts of the world, it is expected that the churner’s patterns would differ as well. The US telecom market has been addressed by [9,10, 11], Asian markets have been in the focus of [3,6,7,19], and South America (Brazil) is the market analyzed by [4]. Other researchers do not discover the location of the operator [14,20], or use publicly available datasets [10,18]. There is a visible lack of papers focusing on the European market (except [13]).

Last, but not least, we would like to mention the issue of defining prepaid churn, or the lack of a clear definition of a churned prepaid customer. To the best knowledge of the authors only [13] defines in detail which prepaid customers are marked as churned. We find the outcome definition to be of high importance and we address this issue in depth in Section 4.2 and 4.5, where we also propose our own definition(s).

3 Customer Experience Management

To put our churn management activities in perspective, we see them as part of a wider ranging company effort to manage and optimize the customer experience. Products and services (e.g. telecommunication services) tend to become commodities over time; therefore, managing the customers' experience can be a source of sustainable competitive advantage [21]. Pine and Gilmore [21] state that experiences are as distinct from services as services are from goods, thus arguing that authentic experiences are the next evolutionary step in creating customer value. Smith and Wheeler claim that branded customer experience drives customer loyalty and profitability [22]. Customer Experience Management (CEM) is the process of strategically managing and optimizing these experiences across all customer touch-points and channels, in interactions that are either customer initiated or company driven, direct or intermediated [23,24].

CEM is closely related to its predecessor Customer Relationship Management (CRM). The distinction is somewhat artificial, but one could argue that CRM focuses more on providing a 360 degrees view of the current relationship and managing customer processes efficiently, whereas CEM provides tools and methodologies to understand, improve and extend the relationship by optimizing the overall customer experience. For mobile telecommunication providers (and many other businesses) the key areas within an overall CEM strategy are customer acquisition, revenue stimulation for existing customers and customer retention management (churn management). Churn models, in particular, can be important components of overall strategies that make these predictions actionable to drive intelligent interactions and experiences.

3.1 Measuring the Customer Experience

The focus of this paper is on building better prepaid churn models, both from a predictive power perspective, as well as from the point of view of predicting behavior that is actionable and makes business sense. For instance, predicting short term inactivity for a base that includes many already dormant accounts will result in a powerful, yet not very useful, model. Embedding the churn models in larger strategies to optimize experience in real time is more the scope of future research. From a CEM perspective, our topic of interest is measuring past customer experience as a potential input to churn models.

We have created a theoretical business framework for measuring customer experience in mobile telecommunications in ideal circumstances (Figure 1). To the best knowledge of the authors, this is a first attempt to create a framework of this kind and purpose. This framework provides a conceptual roadmap and direction for the coming years for the mobile telecom operator where this research was performed. In the short term, it is very challenging to measure most of these dimensions reliably. The framework contains three levels: Aspects, KPIs and Data Sources. The Aspects level represents the various aspects of customer experience in mobile telecommunications. The KPIs level defines possible measurements for the different aspects of customer experience. The Data Source level describes the potential location of various KPIs. This framework has been designed for mobile telecommunications, but it can easily be customized for other business to consumer industries by replacing the industry specific aspects (Handset, Network Usage and Network Usability).

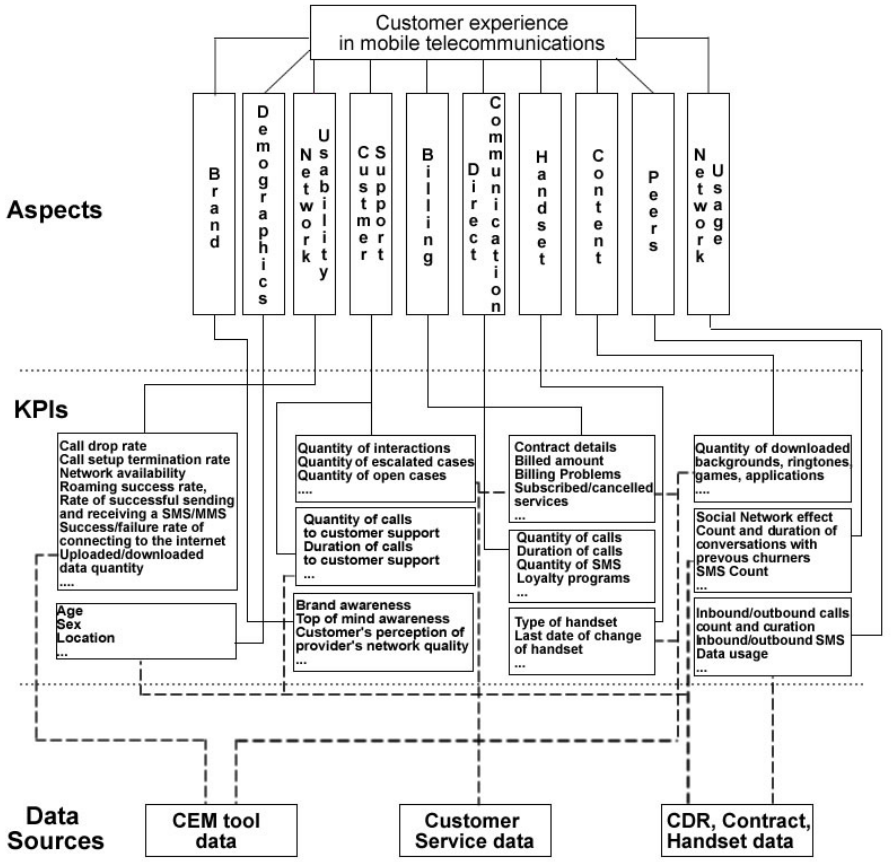
Customer experience in mobile telecommunications
Aspects
Brand
Demographics
Network Usability
Customer Support
Billing
Direct Communication
Handset
Content
Peers
Network Usage
KPIs
Call drop rate
Call setup termination rate
Network availability
Roaming success rate,
Rate of successful sending and receiving a SMS/MMS
Success/failure rate of connecting to the internet
Uploaded/downloaded data quantity
....
Age
Sex
Location
...
Quantity of interactions
Quantity of escalated cases
Quantity of open cases
....
Quantity of calls to customer support
Duration of calls to customer support
...
Brand awareness
Top of mind awareness
Customer's perception of provider's network quality
...
Contract details
Billed amount
Billing Problems
Subscribed/cancelled services
...
Quantity of calls
Duration of calls
Quantity of SMS
Loyalty programs
...
Type of handset
Last date of change of handset
...
Quantity of downloaded backgrounds, ringtones, games, applications
....
Social Network effect
Count and duration of conversations with prevous churners
SMS Count
...
Inbound/outbound calls count and curation
Inbound/outbound SMS
Data usage
...
Data Sources
CEM tool data
Customer Service data
CDR, Contract, Handset data

Fig. 1. Customer Experience Framework for Mobile Telecommunications

We are interested in the relative value of the CEM KPIs for general marketing use, as well as the business value of the software tools used to collect and analyze them. Specifically for prepaid churn prediction, the following six aspects of the CEM framework presented above can represent this added value, as they have not been used before for prepaid churn prediction: Brand, Network Usability, Customer Support, Direct Communication and Content.

Network usage, Billing and Handset have been addressed by [12,13], while Peers (Social Networks) is the topic of [14,19,20]. The Demographics aspect is not applicable to prepaid subscribers, because for most of them this data is not available.

For this paper we want to particularly focus on the added value of the CEM tool data source. This tool measures service quality (dropped calls etc.) for operational purposes, but claims are made that this data is also of key importance for prepaid churn management. We want to validate whether these claims are warranted from a business case perspective, assuming that traditional usage and customer relationship data is already available.

4 Research Setup

In order to examine the influence of the CEM metrics, variations in population sample and outcome definitions, we constructed three separate experiments which are described in the subsections below.

A set of general rules applies to all three. The predictors were recorded at the end of March 2009 with one, three and six months of history. The random sample of around 10% of the prepaid customer base was divided into training, validation and test sets, using the 50:25:25 ratio. It was not necessary to oversample on churn, due to the choice of the software tool, which is able to handle unbalanced distributions of the outcome. The validation set was not used for developing the models in any automated way (e.g. model parameter setting or pruning using automated cross validation). It was used as an additional test set and for manual verification of the univariate performance of variables on an independent dataset to avoid overfitting. In other words, from an algorithm perspective it could have been called a test set, but from a methodological perspective we want to take a cautious approach and call it a validation set, because in theory the analyst himself could overfit the data. The analyst does not have access to test set results in data preparation.

In our situation hold out validation rather than *n*-fold cross validation was both sufficient and more practical. If test datasets available are large enough, a simple holdout provides a true indication of future performance (see also [16]). Note we had ten times more instances available in the source data set. Hold out validation is also more practical given that it results in a model along with validation and test performance estimates. Cross validation at best results in a choice of an optimal modeling approach rather than a model. Hence, one would still need to build a single model, and likely require hold out validation to estimate the test set performance for the resulting model, because it might differ from the cross validation performance.

The next section will provide more technical details on the end to end data mining process followed.

4.1 End to End Data Mining Process

The commercial data mining tool Predictive Analytics Director [17] was used for automated data preparation (attribute discretization, grouping and selection), modeling (logistic regression, decision trees) and model evaluation. All the steps in the data mining process, including the data preparation step, are actually decoupled from each other, with the goal of making the process more manageable and providing a factory approach to model building. On average, we do not expect this to have a negative impact on model performance or robustness. For instance, supervised discretization of continuous variables prior to modeling can sometimes even improve the accuracy of the model [25]. As another example, wrapper based approaches for predictor selection (predictor selection and modeling integrated) are not inherently better than filter based approaches (predictor selection separate from modeling step) [26].

The objective of data analysis is to transform the various attributes, a.k.a. variables (columns in the flat table), and identify the most informative ones among them at this stage, from a univariate perspective. A variable is considered informative if it has a certain level of influence over the outcome variable (which in this case is churn). Both statistical and graphical tools are used to establish this degree of influence.

In this research, the chosen statistical criterion for evaluating the predictive performance of both variables and models with relation to churn is the Coefficient of Concordance (henceforth CoC), which is a rank order correlation measure, equivalent to Area under the ROC (AUC) [17]. The major benefit of rank order correlation measures compared to basic measures such as accuracy is that these measures are not sensitive to skewed distributions of the outcome. For instance, assume a binary target for which one of the outcomes is very rare, e.g. churn. A majority vote model, which in practice is useless for selecting prospective churners, would have accuracy equal to the percentage of the majority class, i.e. 1-churn_rate (e.g. if churn_rate=1%, accuracy=99%). However, in such a case, the values of CoC or AUC would only be 50 or 50%, respectively (equal to the values of random choice), which is the lowest value possible for real world models, as the prediction does not provide useful information to rank instances.

The data analysis process begins by discretizing continuous variables into a large number of bins. Numeric bins and symbolic values without statistically significant difference in churn rate are then grouped together. This balances the training set performance (many groups with varying churn levels) with out-of-sample (validation and test set) and out-of-time robustness (as many instances per group to provide robust estimates of churn for a group). Basically, this is a supervised, bottom-up approach to discretization of continuous variables and merging of numeric bins and symbolic values into groups that display a significant difference in the outcome. The analyst can inspect the resulting histograms, and optionally change the parameters of the process, for instance the significance thresholds for merging bins and symbols.

After the initial data analysis, for the purpose of predictor selection, an automated procedure is used to group the variables that are correlated, independent of the target. A given predictor may have a high univariate performance, but also be correlated with other candidate predictors that are even stronger, hence not adding value to a model (subset predictor selection rather than univariate predictor selection). The user has three options when selecting predictors to be used for modeling: all predictors, only the best of each group, or manual selection. In our case we used the best predictors of each group as a starting point, and then experimented with minimal further manual selection, for instance by removing predictors from the bottom scoring groups altogether, or changing the parameters of the automated grouping process to force more or less groups. The main rationale for this was to force the use of CEM variables in order to enable a comparison.

The full dataset consisted of more than 700 attributes, and the optimal number of predictors for final models typically ranges between 5 and 10 predictors. These volumes of attributes in the raw data versus the final model are quite typical for real world data mining, at least for marketing purposes; the full modeling table is reused as a basis for multiple modeling purposes, as it aims to capture all aspects relevant to customer experience (see also Figure 1). See the results section (Section 5) for more information of the types of predictors selected.

For logistic regression models, the raw predictor values are replaced with a normalized rank score concordant with the churn rate for the group (e.g. the predictor ‘age’ is divided in various age ranges; if the group of 22-26 years shows higher churn rate on the training set, a higher score will be assigned to it). Logistic regression models are then built on the recoded data, to allow capturing of complex non-linear behavior, whilst using a stable low-variance modeler. Note this requires the groups to contain a significant number of cases to prevent ‘leakage’ of the outcome back to the predictors. In our implementation, decision trees were forced to split between groups only, i.e. the grouped bins of discretized variables, thus not on the raw data (supervised discretization prior to induction [25]). Our motivation for this was to provide a level playing field for both algorithms (logistic regression and decision trees) and to ensure that only the ‘analyst approved’ discretization was used. We used the CHAID splitting criterion, which selects the split points based on statistical significance as measured by the Chi Squared statistic [27]. This criterion allows merging of both adjacent and non-adjacent bins of a discretized variable when making the splits. For example, if a split on variable ‘age’ is performed, and the bins 22-26 and 30-34 display similar level of significance, only one split for ‘age’ will be created, containing both these intervals (i.e. age in 22-26,30-34 and age not in 22-26,30-34). The end result of our procedure is a binary tree. We are aware that CHAID and other decision tree induction methods are capable of producing n-ary trees [28], which could have a somewhat better performance. However, it was not our goal to test algorithm performance, thus we used a standard binary decision tree.

The modeling process results in scoring models: a rank score concordant with the probability of being a churner is allocated to each of the instances. The CoC (AUC) measure is used to measure model quality. In addition, we use gain charts as visual representation of model performance. On the y-axis, these charts show the captured proportion of the desired class (i.e. churners in selection divided by total number of churners) with increasing selection sizes (x-axis, from highest scoring to lowest scoring) (see Figure 3).

4.2 Definition of Prepaid Churn and Initial Sample

Defining prepaid churn is more complex than defining postpaid churn, as there is no concept of contract termination. Prepaid customers are deemed as churned if they are no longer ‘active’ for a certain period of time.

The working definition of ‘activity’ within the company where this research was conducted was used as operational definition. This definition is based on various aspects of usage of telecom services. On one hand, it is very detailed and captures the various aspects of activity within mobile telecom industry. On the other hand, it simplifies the data acquisition.

Definition 1: Operational definition of activity: Outgoing (initiated) calls, top-up (recharge) of the account, sent SMS/MMS messages and received (incoming) and answered calls are considered as an activity. Received calls without picking up, received SMS/MMS messages and bonus credit top ups awarded by the company are NOT considered as an activity.

Involuntary prepaid churn in this company occurs after six consecutive months of no activity. After this, the customer can no longer use the services of the company via that particular SIM card, and the phone number may be reissued to another client after a certain period. However, six months of no activity is too long of a period to investigate voluntary prepaid churn. Most of the internal projects investigating prepaid churn within this company consider customers to have churned if they had not been active between two and three consecutive months. Therefore, we propose the following

Definition 2: Operational Prepaid Churn Definition: Customers are marked to have churned if they are registered to have two consecutive months of no activity, or more.

It was also necessary to set certain boundaries for the population taken into account for the modeling process. The population boundaries are set in Definition 3.

Definition 3: Population definition: The population consists of prepaid subscribers that meet the following two conditions:

1. They have at least one registered activity in the last 15 days.
2. Their first activity date is at least four months before.

The first condition enables avoiding subscribers who can already be classified as churners using the definition above. Arguably, to avoid this group, it would be sufficient to set the limit in condition 1 to 59 days, but this would make the prediction task trivial. The boundary was set to 15 days as a balance between excessive reduction of the sample and limitation of the information spillover (the subscribers inactive for 30 days have already begun to display churn behavior). The second condition is useful for avoiding frequent churners (it implies loyalty) and for avoiding bias when measuring communication with previous churners.

In summary, for the particular data set constructed, all predictors were measured in the first trimester of 2009; churn was measured two months later; inactivity at recording was limited to maximum 15 days and first activity date had to be minimum four months prior to recording. This served as baseline data set.

4.3 Experiment A: Addition of CEM Parameters

Capturing the service quality oriented CEM metrics was performed by using a CEM tool deployed in the company. This tool suffered from two limitations. First, it contained only 40 days of user history. In order to give the CEM parameters a fair competing chance, we only used traditional parameters with one month history.

Second, this tool did not cover the entire customer experience as depicted on the CEM Framework on Figure 1. It is more of a network experience measurement tool due to the fact that it is focused on the more hygienic aspects of customer experience (aspects noticed by customers only if they are absent or have low quality). This database contains only the Network Usability aspect (e.g. Call Success Rate, SMS Success Rate, Call Setup Duration, etc.) and the Content aspect (e.g. count of accessing company's website for downloading ringtones, backgrounds, etc.) of our CEM framework.

However, this tool did not capture any data related to Customer Support; therefore, Customer Support data was added from the CDRs. Other aspects of the CEM framework, such as Network Usage, Handset and Peers, which were recorded from the company's CDRs and customer subscription data, were used for benchmarking purposes (to test the added value of parameters that are viewed exclusively as CEM KPIs). Therefore, they cannot be treated as potential contributions of CEM to the model's performance, but rather as traditional parameters as described in Section 2. Demographic information was not available for prepaid subscribers.

The two remaining aspects, Brand and Direct Communication, could not have been analyzed because the company did not have data appropriate for data mining (Brand), or did not communicate proactively to its prepaid customers. Therefore, only three aspects were left for analysis: Network Usability, Content and Customer Support. Hence, the only added value stemming from using CEM KPIs on prepaid churn modeling in this research can exclusively originate from one of these three aspects of the CEM framework. In order to determine the added value of CEM, we compare models consisting only of "traditional" parameters, to models consisting of both "traditional" and CEM parameters.

4.4 Experiment B: Variations in Population Sample

In order to test impact of the variations in the population sample we changed the activity restriction into maximum 30 and 0 days of inactivity; therefore, we change condition 1 from definition 3 into condition 1a and condition 1b, respectively.

Condition 1a: Subscribers must have at least one activity in the last 30 days.
Condition 1b: Subscribers must have at least one activity in the last day before recording.

The reasons for varying the sample on maximum period of allowed inactivity are threefold. First, our intention was to inspect the impact of these changes on the models' performance. Second, we wanted to test the contribution of the CEM parameters under different circumstances. Third, the typical period of users' inactivity, before they become unavailable for contact and retention, is disputable. Additional motivation for experiment B can also be found in the results of experiment A.

Furthermore, using zero days of inactivity can be seen as a way to avoid information spillover. We defined churn as uninterrupted inactivity for two months. It is clear that the best predictor of this is if a user has already been inactive for a certain period. In other words, users have already started to display churn behavior. This spillover cannot happen when the inactivity period at recording is limited to zero. Additionally, these subscribers are likely to still be available for contacting. We can expect that churn is harder to predict for this subgroup of subscribers.

4.5 Experiment C: Change in Outcome Definition

Since there is no general consensus on a practical definition of prepaid churn in mobile telecommunications, it is reasonable to experiment with different definitions. In practical terms, this is the first question that arises during a prepaid churn modeling project. A change in the churn definition not only affects the churn rate, but also the future retainability of prepaid consumers deemed as churners using that definition.

Our motivation for changing the outcome definition was also to investigate the impact of this change on the performance of the models, while keeping the same sample and dataset, and test the added value of the CEM parameters in this situation. For these reasons, we introduce a so-called "grace" period of 15 days, thus changing Definition 2 into

Definition 2b: Customers are marked to have churned if they are registered to have at least one activity in the first 15 days after recording, followed by 2 consecutive months of no activity, or more.

This outcome definition also resolves the information spillover issue discussed in Section 4.2. Namely, subscribers must have an activity in the first 15 days ("grace period") after recording, thus breaking out of their already displayed churn behavior. Subscribers inactive in this grace period, who continue to be inactive in the future, are labeled as non-churners, because they have already churned at the time of recording, and are of no interest. In addition, subscribers recognized as future churners using this definition are more likely to be available for contacting and retention, because we are aiming at subscribers who are first active for 15 days, and then inactive for two months or more.

5 Results

In this section we report the results of the experiments of adding CEM data, changing the population based on inactivity duration and changing the outcome definition, or experiments A, B and C, respectively (see Table 1 for highlights). As explained in Section 4, we are using CoC as the main quantitative measure to compare models. In addition we provide gain charts to provide a visual reference for the performance of the models on the training set. These charts are used for illustrative purposes only, namely to visualize CoC performance through percentage of detected churners in a given percentage of the population scores (e.g. 78% of churners within top 20% of population scores).

It is worth mentioning that performance-wise, two models can be compared only if they have been created under the same experimental setup (e.g. Models A_Excl_CEM and A_Incl_CEM can be compared, Models A_Excl_CEM and C_Excl_CEM cannot).

During experiment A, it became immediately clear that certain aspects of the CEM framework will not be able to add value in the case of prepaid churn prediction. Namely, a very small percentage of the prepaid users had contact with Customer Service, or had downloaded any content from the provider's website. Therefore, none of the variables from these two aspects of the CEM Framework presented on Figure 1 could have been a good predictor. Furthermore, a very small percentage of these users used data services. Hence, Network Usability parameters related to mobile internet usage were also not good predictors. The only CEM predictors left to add value were the voice call and SMS related KPIs from the Network Usability of the CEM Framework.

Model A_Excl_CEM contains only traditional parameters and is built on six predictors, containing only one month history. Adding parameters with longer history did not change the performance of the model substantially. The strongest predictor for churn was the inactivity period, followed by the remaining credit on the user's prepaid account. Other variables included were the handset and count and duration of calls. Model A_Excl_CEM had the following CoCs: 87.8 on the training set, 85.6 on the validation set, and 87.5 on the test set. Model A_Incl_CEM contains the same six variables, plus two CEM parameters (from the Network Usability Aspect of the CEM Framework), which were selected using a trial and error process based on their respective CoCs and the predictor group to which they belonged. Model A_Incl_CEM had the following CoCs: 87.9 on the training set, 85.6 on the validation set, and 87.5 on the test set. Thus, the results were reasonably consistent throughout the training, validation and testing sets. The only difference in the performances of these models was on the training set, which is not the best measure of model performance. Both these models were built using decision trees, but models built on the same variables using logistic regression performed just as well.

For visual reference only, we present the gain chart of the two models created for experiment A (on the training set) on Figure 3c. The difference in CoC of 0.1 (A_Excl_CEM-87.8 vs. A_Incl_CEM-87.9) is not even visible on the gain chart. Their gain charts are identical up to 50% of the population. Both models were able to identify 78% of the churners within the top 20% of the population scores (a lift of 3.9 in the top 20% of population scores). This is representative of the test set performance as well, because the CoC on the test set of both models (87.5) is very similar.

The inability of the CEM variables to substantially improve the base model is due to two reasons. First, a large number of these variables have a very low CoC (univariate performance). Second, even when the CoC is relatively high, in most groups there are traditional (Non-CEM) predictors with CoC values higher than the CEM variables in the same group. Such is the case in the highest ranked group 1, which is depicted on figure 2.

In order to illustrate the reasons for the weak effect on model performance improvement of the CEM variables, we isolated only the eight variables from model A_Incl_CEM, and ran the automatic grouping operation, with more strict grouping parameter settings than used previously, to force further grouping. The results of this exercise are presented on Table 2. One of the CEM parameters, CEM_var_x, has a reasonably high performance, but is in the same group as three other traditional variables, which means it has a degree of correlation with them, and has the lowest CoC in that group. This explains why CEM_var_x does not improve the model performance substantially. The second CEM variable in this model, CEM_var_y is in a class of its own, but it does not have a very high CoC; therefore, it cannot add substantial value to model performance.

Table 1. Sample size, churn rate and CoCs in experiments A, B1a, B1b and C

Experiment	Inactivity Allowed	Sample size	Churn Rate (%)	Max CoC Train Set	Max CoC Validation Set	Max CoC Test Set
A	15 days	62565	3.88	87.9	85.6	87.5
B1a	30 days	67986	6.09	89.2	89.0	88.7
B1b	0 days	32104	0.70	85.3	77.1	79.5
C (Definition change)	15 days	62565	2.40	72.7	68.6	68.5

Table 2. Grouping of variables of Model A_Incl_CEM

Predictors	CEM Framework Aspect	Group	Performance (CoC)
Inactivity Period	Network Usage	Group 1	81.1
Trad_var_x	Network Usage	Group 2	70.8
Trad_var_y	Network Usage	Group 2	70.6
Trad_var_z	Network Usage	Group 2	69.8
CEM_var_x	Network Usability	Group 2	68.9
Remaining credit	Billing	Group 3	67.4
CEM_var_y	Network Usability	Group 4	63.2
Handset type	Handset	Group 5	57.2

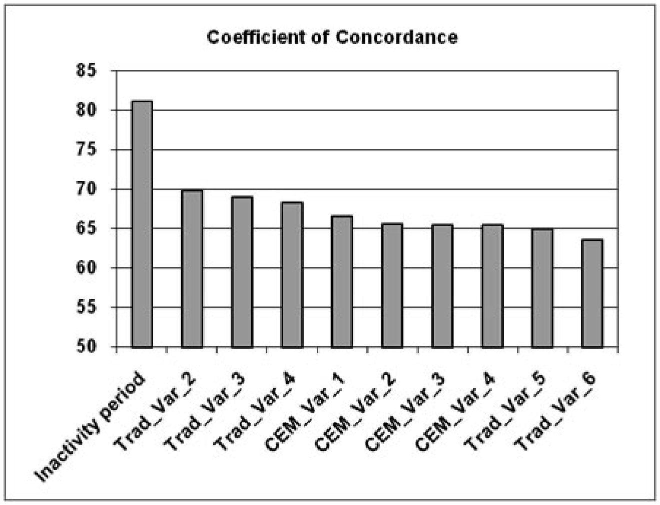

Fig. 2. Coefficient of Concordance of predictors grouped in group 1 for experiment A.

Please note that Table 2 and Figure 2 do not present the same groups of predictors.

Due to the strong influence of the inactivity period in experiment A, we decided to vary the population sample by changing the inactivity limit at recording to 30 and zero days (experiments B1a and B1b, respectively). Model B1a_Excl_CEM was built on seven non-CEM variables, very similar to the ones used in model A_Excl_CEM, using decision trees. B1a_Incl_CEM was built on the same seven variables plus two CEM parameters (again from the Network Usability Aspect of the CEM Framework), selected using a trial and error process based on their respective CoCs and the predictor group to which they belonged, similar to experiment A.

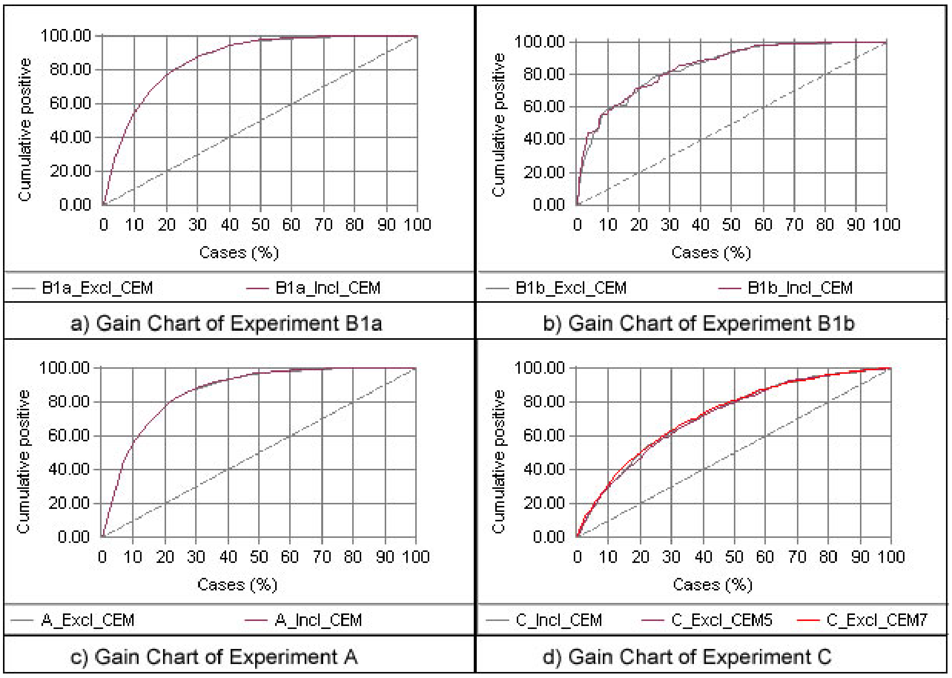

Fig. 3. Gain chart of models for experiment A, B and C (training set).

Both models had the following CoCs of 89.2 on the training set, 89.0 on the validation set and 88.7 on the test set, which is a very stable performance. Once again, there was no visible performance difference (see Figure 3a). The influence of the inactivity period was even stronger, once again followed by the remaining credit. The gain chart on Figure 3a shows that both models identified about 78% of churners in the top 20% of the population scores.

In experiment B1b, the model B1b_Excl_CEM is built on only two non-CEM parameters (prepaid account balance and count of calls), because adding more variables caused overfitting (i.e. higher differences in CoC on the training, validation and test set). Model B1b_Incl_CEM was built on the same two variables, plus one more CEM parameter (call setup duration). This time, both models were built using logistic regression. In this situation, models built on decision trees were less stable across the three datasets, and their performance on the test set was lower, even though the performance on the training set was similar (overfitting). Nevertheless, even when using logistic regression, the models were not as stable as it had been in the previous cases.

Model B1b_Excl_CEM had the following CoCs: 84.9 on the training set, 77.1 on the validation set and 79.5 on the test set. Model B1b_Incl_CEM had the following CoCs: 85.3 on the training set, 74.0 on the validation set and 79.5 on the test set. Conclusively, adding CEM variables did not improve performance. On the contrary there is a performance drop on the validation set. Please note that the gain chart on Figure 3b is somewhat optimistic, because it was built on the training set. The performance on the test set is lower. Nevertheless, both models are able to identify 71% of the churners on the top 20% of the population scores in the training set, which is lower than the models in the previous experiments.

Finally, we present results for experiment C. The altered churn definition included the requirement of activity in the first two weeks of the outcome period, followed by a period of no activity of two months or more. Model C_Excl_CEM5 contained only traditional parameters and was built on five predictors. This model had the following CoCs: 72.2 on the training set, 68.5 on the validation set and 67.9 on the test set. Model C_Incl_CEM contained the same five variables, plus two CEM parameters. The CoCs of that model were 72.7 on the training set, 68.6 on the validation set and 67.9 on the test set. Both models were built using logistic regression, because these had a better performance on the test set when compared to models built on decision trees. There is an insubstantial 0.1 CoC improvement on the validation set, which is also achievable by adding two non-CEM parameters (model C_Excl_CEM7).

The gain charts of these models' performances on the training set are presented on Figure 3d. The maximum churners percentage achieved within the top 20% of population sample here is 50%. Note that this number would be even lower on the test set. Nevertheless, this is almost 30% identified churners less (in the top 20% of population scores) than what was achieved in experiments A and B1a.

The results of experiment C were less consistent throughout the three datasets when compared to experiments A and B1a, but somewhat more consistent when compared to B1b. It is interesting that the inactivity period was *not* a factor in these models, even though 15 days of inactivity were allowed at recording. The most powerful predictors were the remaining credit, the handset and the call count. Similarly to experiment A, the inability of CEM variables to improve the model performance in both experiments B and C is due to either low CoC or correlation with stronger traditional predictors.

6 Discussion and Future Research

We conducted three experiments in order to compare the influence of new CEM parameters, as well as changes in sample population and outcome definition, on prepaid churn model's performance.

CEM is advertised in literature to have an added value in predicting churn, but this was not the case in our experiments. Models without CEM parameters performed almost the same as models which included CEM parameters in all three experiments. However, we only tested the CEM parameters with "hygienic" nature, which are only noticed when absent or unsatisfactory. The softer aspects of CEM remain untested. However, at this point we expect that it will be hard to find new non behavioral predictors with sufficient predictive power compared to the behavioral data. Even though the CEM data we had available contained only 40 days of history, it is unlikely that longer history would change the outcome, because the non-CEM parameters used by the models in most cases also had only one month history.

The rationale behind the low added value of CEM parameters for prepaid churn modeling may be found in several factors.

The prepaid customers themselves are the first factor. Prepaid customers that were subject to this research had average call duration of around one minute. A very low percentage of prepaid users used data services or have called the Customer Service in the research period. In other words, these events are rare, which limits the potential to become an interesting predictor. Next, prepaid users are mostly interested in controlling (reducing) their mobile phone expenses; otherwise, they could switch to using post-paid services that offer less expensive calling tariffs. The interest of prepaid users to control their mobile phone expenses may have been enhanced by the Global Financial Crisis of 2007/2008. To summarize, prepaid customers are more concerned with the quantity of experiences, which is measured by traditional predictors (Section 2) rather than the quality of their experiences, measured by the CEM parameters. The only exception is the handset which is a parameter that deeply influences the quality of the experiences, and is also regarded as a lifestyle product.

The second factor that could explain the low added value of CEM parameters was the high network quality. The percentage of customers experiencing network problems (negative experiences) is very small. However, the network quality cannot be seriously tested on average call duration of one minute.

The third explanation for the low added value of CEM parameters is that the quality parameters are correlated (to a degree) to their quantitative counterparts (e.g. number of dropped calls is correlated to a degree with number of calls).

Changing the population sample by varying the inactivity limit at the time of recording between 15 and 30 days also did not contribute to a substantial change in model performance. Models in experiments A and B1a had CoCs of about 88 and 89, respectfully (they identified between 78% and 79% of churners in the top 20% of the population). However, the churn rate did change drastically (there are almost twice more churners in B1a), while the sample size change was less than 10%, as presented in Table 1. These changes are even more evident at experiment B1b, where no inactivity was allowed at recording. Here, the sample size is twice smaller when compared to experiment A, while the churn rate is five times smaller when compared to experiment A, and even 9 times smaller when compared to experiment B1a. Due to the very low churn rate and the higher complexity of the task, the performance in this experiment was lower. The maximum CoC achieved on the test set was 79.5, which is nine CoC points less when compared to the other two experiments; this results in a lower percentage of churners in the top 20% of the population (8% less when comparing training sets, but the difference is higher on the test set). The change of allowed inactivity period to zero also influenced the model stability (i.e. consistent performance on all three datasets - training, validation and test). Note that we used a very small number of variables (two and three) in this experiment's models, in order to avoid overfitting. Once again it is important to emphasize that customers with zero days of inactivity at the time of deployment are more likely to be available for retention than customers with 15 or 30 days of inactivity.

The most dramatic change in model performance was shown when the outcome definition was changed and a so-called grace period of 15 days was included to mirror the inactivity period allowed at the time of recording. The model performance dropped by 20 CoC points on the test set, compared to experiments A and B1a (results in an almost 30% drop in identified churners in the top 20% percent of the sample on the training set, and even more on the test set). This does not mean that models built under experimental setup C are worse than the others. It merely implies the expected performance under such conditions. This steep decline is due to the complicated churn definition we deployed (we targeted an inactivity-activity-inactivity pattern). In this case, the inactivity period, that was a dominating variable in experiments A and B1a, was not a factor at all. The benefit of using such a definition is that upon deployment, the identified churners are likely to have an activity in the next 15 days, which makes them available for retention.

The focus of the research was on the impact of the experimental setup, rather than the algorithm used to create the model. Therefore, we used standard data mining algorithms, such as decision trees and logistic regression. Having said that, we would like to emphasize that in experiments A and B1a, there was barely any difference on model performance that could be attributed to the usage of the particular algorithms. In experiments B1b and C, there was a small difference in performance of algorithms, but it was not as substantial as changing the outcome definition or changing the allowed period of inactivity at recording to zero. In these two experiments, logistic regression had a more stable performance on the training, validation and test sets, compared to decision trees.

In terms of directions for future research, we would like to investigate a richer set of customer experience data, particularly around proactive communications and brand. Additionally, it would be worthwhile to investigate the relations between duration of inactivity and availability of subscribers for retention, by inspecting their presence on the network (regardless of making calls). Last but not least, we would like to take into account the feedback from retention campaigns, in order to focus on "retainable churners." After all, the end target of churn prediction is retention.

7 Conclusion

In this paper, we presented how performance of prepaid churn models changes when varying the conditions in three different dimensions: data- by adding CEM parameters; population sample- by limiting the inactivity period at the time of recording to 15, 30 and zero days, respectively; and outcome definition- by introducing a so-called grace period of 15 days after the time of recording, in which customers must make an activity in order to be classified as churners.

Adding the CEM parameters into the models did not add substantial value in model performance under any of these conditions. Similarly, switching the population sample on the period of inactivity at the time of recording between 15 and 30 days did not influence model performance, only the sample size and churn rate. When we changed the population sample by disallowing inactivity at time of recording, apart from the change in sample size and churn rate there was also a drop in performance and stability of the models. However, this drop in performance was not nearly as high as the one that occurred when changing the outcome definition by setting a grace period, thus making the behavior to be predicted more complex. This change obviously influenced the churn rate as well. Nevertheless, the latter two approaches should provide more time for retention.

References

1. Hung, S., Yen, D. C., Wang, H.: Applying data mining to telecom churn management. Expert Systems with Applications 31(3), 515-524 (2006)
2. Mozer, M., Wolniewicz, R., Johnson, E., Kaushansky, H.: Churn reduction in the wireless industry. Advances in Neural Information Processing Systems 12, 935-941, MIT Press, Cambridge (2000)
3. Kim, H., Yoon, C.: Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy 28(9-10), 751-765 (2004)
4. Ferreira, J., Vellasco, M., Pachecco, M., Barbosa, C.: Data mining techniques on the evaluation of wireless churn. ESANN' 2004 proceedings - European Symposium on Artificial Neural Networks, pp. 483-488, Bruges (2004)
5. Neslin, S., Gupta, S., Kamakura, W., Lu, J., Mason, C.: Detection defection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing research 43(2), 204-211 (2006)
6. Au, W., Chan, K., Yao, X.: A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation 7(6), 532-545 (2003)
7. Hwang, H., Jung, T., Suh, E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26, 181-188 (2004)
8. Wei, C., Chiu, I.: Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications 23, 103-112 (2002)
9. Lemmens, A., Croux, C.: Bagging and Boosting Classification Trees to Predict Churn, Journal of Marketing Research 43 (2), 276-286 (2006)
10. Lima, E., Mues, C., Baesens, B.: Domain knowledge integration in data mining using decision tables: Case studies in churn prediction. Journal of the Operational Research Society 60, 1096-1106 (2009)
11. Datta, P., Masand, B., Mani, D., Li, B.: Automated cellular modeling and prediction on a large scale. Artificial Intelligence Review 14, 485-502 (2000)
12. Archaux, C., Martin, A., Khenchaf, A.: An SVM based churn detector in prepaid mobile telephony. International Conference on Information & Communication Technologies (ICTTA), pp. 19-23, Damascus (2004)
13. Alberts, L. J. S. M.: Churn Prediction in the Mobile Telecommunications Industry, MSc Thesis, Department of General Sciences, Maastricht University (2006)
14. Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A.: Social Ties and their Relevance to Churn in Mobile Telecom Networks. Proceedings of the 11th international conference on Extending database technology, pp. 668-677, Nantes (2008)
15. Radosavljevik, D., van der Putten, P., Kyllesbech Larsen, K.: The Impact of Experimental Setup on Prepaid Churn Modeling: Data, Population and Outcome Definition. In:

Bichindaritz, I., Perner, P. Ruß, G. (Eds.), Advances in Data Mining, Workshop Proceedings, pp. 14-27, IBaI Publishing, Leipzig (2010)

16. Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. (2nd ed.) Morgan Kaufmann, San Francisco (2005)
17. Chordiant Software: Chordiant Predictive Analytics Director [Computer software]. Chordiant Software, now part of Pegasystems (www.pega.com) (2008)
18. Verbeke, W., Martens, D., Mues, C., Baesens, B.: Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications (2010), doi:10.1016/j.eswa.2010.08.023
19. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1039-1048, Washington, DC (2010)
20. Richter, Y., Yom-Tov, E., Slonim, N.: Predicting customer churn in mobile networks through analysis of social groups, Proceedings of the SIAM International Conference on Data Mining, pp. 732-741, Columbus (2010)
21. Pine, B.J.I., Gilmore, J.H.: The Experience Economy. Harvard Business School Press, Boston (1999)
22. Smith, S.. Wheeler, J.: Managing the Customer Experience: Turning Customers into Advocates. FT Prentice Hall, Harlow (2002)
23. Meyer, C., Schwager, A.: Understanding Customer Experience. Harvard Business Review 85 (2), 116-126 (2007)
24. Schmitt, B.H.: Customer experience management: A revolutionary approach to connecting with your customers. John Wiley & Sons, Hoboken (2003)
25. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Prieditis, A., Russell., S (Eds.): Proceedings of the Twelfth International Conference on Machine Learning, pp. 194-202 Morgan Kaufmann, San Francisco (1995)
26. Tsamardinos, I., Aliferis, C.: Towards principled feature selection: Relevancy, filters and wrappers. In: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West (2003)
27. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119-127 (1980)
28. Perner, P.: Data Mining on Multimedia Data. LNCS, vol. 2558, Springer Verlag, Berlin (2002)