Incorporating time-dependent covariates into BG-NBD model for churn prediction in non-contractual settings

Andrey Babkin[a], Ina Goldberg[b]

[a] *Andrey Babkin is a Data Scientist at SteppeChange LLC*
andrey@steppechange.com
[b] *Ina Goldberg is a Head of Data Science at the SteppeChange LLC*
ina.goldberg@steppechange.com
address: 3085 Alexis Dr Palo Alto, California 94304-1304 United States
http://steppechange.com

Abstract

Nowadays, churn prediction models in non-contractual settings are gaining increasing interest. In a non-contractual setting the exact moment of customers dropout is unknown. The popular approach to identify active customers is to fit parametric probability model and then infer the probability of being alive from the model and customer's datum. This approach is employed in extension to NBD model (Donald G. Morrison 1988), Pareto/NDB model (David C. Schmittlein 1987) or BG/NBD model (Fader et al. 2005). But despite the respect these models were earned, they can't utilize time-dependent covariates apart from recency and frequency. However, in real-life settings, many other time-dependent covariates are available, for example seasonality or scheduled promotional events. We developed the extension of BG/NBD model which is able to utilize any kind of covariates, including time-dependent variables and monetary values from transactions. Proposed model demonstrated improvements of churn prediction in comparison with BG/NBD model on a real dataset.

Preprint submitted to JOURNAL OF INTERACTIVE MARKETINGNovember 27, 2016

Keywords: Churn prediction, Customer-base analysis, BG-NBD, Non-contractual

Introduction

In general, there are 2 main approaches to non-contractual churn problem: Probability models and Data Mining models.

Probability models use parametric distributions to model customers behavior and find optimal parameters through maximum likelihood approach. This approach is based on articles (Ehrenberg 1959) and (David C. Schmittlein 1987) and developed further in (Fader et al. 2005) and (Fader and Hardie 2009). These authors developed Pareto/NDB and BG/NBD models that predict future customers behavior as well as probability of being alive from past historical transaction data. However, the main drawback of these models is that only recency and frequency data and time-invariant covariates (Fader and Hardie 2007) could be utilized for prediction purposes. This restriction limits the range of predictors that are available in real business settings, for example calendar information could not be used.

In contrast, Data Mining approach is based on applying supervised classification techniques to find probability of churn conditional on past historical transaction data. This approach was developed and reviewed by many authors, for example (Coussement and den Poel 2009), (Jahromi et al. 2010), (Bock and den Poel 2011), (Yu et al. 2011). The comparison of various methods, including Pareto/NBD (probabilistic approach) and several machine learning methods (Data Mining approach) in terms of dropout prediction is provided in (Tamaddoni et al. 2016). The advantage of Data Mining approach is that all available covariates could be utilized for improving churn prediction. However, supervised classification requires target (the fact of customer's attrition) to be available. This is problematic in non-contractual settings as there is no dropout indicator. Instead, data mining approach uses empirically defined targets, for example unusually long transaction vacancy of other judgemental target. Unsupervised methods, for example clustering, were employed in (Jahromi et al. 2010) to define targets for supervised classification. However, the question about relation of obtained clusters and customers who are going to drop out is open.

Our approach synthesizes probability and data mining approaches, that allows to benefit from their advantages and overcome disadvantages. The main idea is to build probabilistic model that includes probability of churn conditional on the past history, then build the sequence of supervised classification tasks that converges to the maximum likelihood of the model.

Problem definition

We follow the definition of the problem as in (Fader et al. 2005):

1. Customer is observed from the beginning of his history.
2. Customer's transactions are traced until the end of period [0,T].
3. The exact moment of churn isn't known.
4. The cohort consists of customers who joined the study during some subset of the observed period (for example in the first 3 months of the period).

The goal is to determine the probability that customer remains active after the end of observations T from his transaction history. Apart from that conditions, transaction time is defined on daily level only. Therefore, we assume that customer makes no more than 1 transaction per day. If there are several transactions during 1 day, we simply aggregate customer's transactions on daily level by summing monetary values and the number of items. Obviously, for different business settings daily granularity could be replaced to hourly granularity ans so on in dependence of available data.

Modelling approach

In this section we will define modelling approach and assumptions that are necessary for the derivation of likelihood and, consequently, for estimation of model parameters. Apart from that, we will highlight aspects of our approach that are different from original BG/NBD model. Our assumptions are very similar to BG/NBD model (Fader et al. 2005), except for that probability of dropout at the end of every transaction is conditioned on the past history and time-dependent covariates that are available by the end of the transaction:

1. As in (Fader et al. 2005): The time between transactions is distributed exponential with transaction rate λ.
2. As in (Fader et al. 2005): Heterogeneity in transaction rates across customers follows a gamma distribution with shape parameter r and scale parameter α:

$$f(\lambda \mid \alpha, r) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}, \lambda > 0 \tag{1}$$

3. After transaction as the moment $t_{i,j}$, a customer becomes inactive with probability p. The probability depends on past history, time-dependent

covariates and churn model parameters:

$$p(dropout \mid \bar{D}_{i,j}, \Theta)$$

where

$$\bar{D}_{i,j} = (t_{i,1}, \ldots, t_{i,x_i}, H_{i,1}, \ldots, H_{i,x_i}) \tag{2}$$

is the history of customer's transactions and time-dependent covariates prior to the moment $t_{i,j}$. The Θ is the set of dropout model's parameters to be estimated.

The assumption 3 is the key distinction from original BG/NBD model. In contrast with being unobservable constant for any given customer, probability of dropout in proposed model depends on history of past transactions and time-dependent covariates that are available by the end of every transaction. This dependence is realized via conditional probability $p(dropout \mid \bar{D}_{i,j}, \Theta)$. Later, we provide the method of estimation for this probability via reduction to the sequence of weighted classification problems.

Derivation of the Likelihood Function on individual level

In this section we derive the likelihood of customer's datum with respect to unobservable parameter λ. This likelihood is needed to derive full likelihood and probability of churn. The derivation of the likelihood on customers level is similar to (Fader et al. 2005) except for terms related to conditional probability of dropout at the end of every transaction.

Consider a customer i who had x_i transactions in the period (0, T] with the transactions occurring at $t_{i,1} \ldots t_{i,x_i}$ and corresponding covariates $H_{i,1} \ldots H_{i,x_i}$:

$$0 \longrightarrow (t_{i,1}, H_{i,1}) \longrightarrow (t_{i,2}, H_{i,2}) \longrightarrow (t_{i,3}, H_{i,3}) \cdots \longrightarrow (t_{i,x_i}, H_{i,x_i}) \longrightarrow T$$

1. As in (Fader et al. 2005), the likelihood of the first transaction occurring at $t_{i,1}$ is a standard exponential likelihood component, which equals

$$\lambda e^{-\lambda t_{i,1}}$$

2. As in (Fader et al. 2005), the likelihood of the j-th transaction occurring at $t_{i,j}$ is the probability of remaining active at $t_{i,j-1}$ times the standard exponential likelihood component, which equals

$$\begin{aligned} &(1 - p_{i,j-1})\lambda e^{-\lambda(t_{i,j} - t_{i,j-1})} \\ &p_{i,j-1} = p(dropout \mid \bar{D}_{i,j-1}, \Theta) \end{aligned} \tag{3}$$

   where $p(dropout \mid \bar{D}_{i,j-1}, \Theta)$ is the probability of churn conditional on model parameters and history prior to $t_{i,j-1}$

3. The likelihood of observing zero purchases in $(t_{i,x_i}, T]$ is the probability the customer became inactive at t_{i,x_i}, plus the probability he remained active but made no purchases in this interval, which equals

$$\begin{aligned} &p_{i,x_i} + (1 - p_{i,x_i})e^{-\lambda c_i} \\ &c_i = T - t_{i,x_i} \end{aligned} \tag{4}$$

Therefore customer-level likelihood is:

$$\begin{aligned} &L(\lambda, \Theta \mid \bar{D}_i) = \prod_{j=1}^{x_i - 1}(1 - p_{i,j})\lambda^{x_i} e^{-\lambda t_{i,x_i}}(p_{i,x_i} + (1 - p_{i,x_i})e^{-\lambda c_i}) \\ &\bar{D}_i = (t_{i,1} \dots t_{i,x_i}, c_i, H_{i,1} \dots H_{i,x_i}) \end{aligned} \tag{5}$$

where $\bar{D}_i$ is the data available for customer i, including both history and the c_i period from last transaction to the end of observations. The main difference between $\bar{D}_i$ and $\bar{D}_{i,j}$ is that the latest term includes only history and covariates that are available prior to $t_{i,j}$, whereas $\bar{D}_i$ includes all the information for customer i, including the distance from the last transaction to the end of observations.

Hereinafter,for simplicity, we denote $L_i(\lambda, \Theta) = L(\lambda, \Theta \mid \bar{D}_i)$

Sample likelihood

In this section we derive expectation of customer's likelihood over unobservable parameter λ. This expectation is needed for parameter estimation as well as for the probability of churn. We derive likelihood with prior from equation (1) as the expectation of $L_i(\lambda, \Theta)$ over λ with respect to (1):

$$L(\alpha, r, \Theta \mid \bar{D}_i) = \int_0^\infty L_i(\lambda, \Theta) \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)} \, d\lambda \tag{6}$$

Substitution of (5) into (6) yields:

$$\begin{aligned}
&L(\alpha, r, \Theta \mid \bar{D}_i) = a(\alpha, r, \bar{D}_i) \left(\prod_{j=1}^{x_i - 1} (1 - p_{i,j}) \right) (p_{i,x_i} * b(\alpha, r, \bar{D}_i) + (1 - p_{i,x_i})) \\
&a(\alpha, r, \bar{D}_i) = \alpha^r \frac{\Gamma(x_i + r)}{\Gamma(r)} \left(t_{i,x_i} + \alpha + c_i \right)^{-(x_i + r)} \\
&b(\alpha, r, \bar{D}_i) = \left(1 + \frac{c_i}{t_{i,x_i} + \alpha} \right)^{-(x_i + r)}
\end{aligned} \tag{7}$$

where x_i is the number of transactions of customer i within the observed period and the $p_{i,j}$ is defined in equation (3).The details about derivation of (7) could be found in Appendix A.

Hereinafter,for simplicity, we denote

$$\begin{aligned}
L_i(\alpha, r, \Theta) &= L(\alpha, r, \Theta \mid \bar{D}_i) \\
a_i(\alpha, r) &= a(\alpha, r, \bar{D}_i) \\
b_i(\alpha, r) &= b(\alpha, r, \bar{D}_i)
\end{aligned} \tag{8}$$

Therefore, individual log likelihood is:

$$\log L_i(\alpha, r, \Theta) = \log a_i(\alpha, r) + \Big(\sum_{j=1}^{x_i - 1} \log\left(1 - p_{i,j}\right) \Big) + \log\left(p_{i,x_i} * b_i(\alpha, r) + (1 - p_{i,x_i})\right) \tag{9}$$

Full log likelihood of the given sample is:

$$LL(\alpha, r, \Theta) = \sum_{i=1}^{N} \log L_i(\alpha, r, \Theta)$$

84 Probability of churn

In this section we derive the key result of the model which is the probability of churn for given customer after the end of observed period. The way of deriving the formula is similar to (Fader and Hardie 2008). Lets denote I as indicator of the churn event. Therefore, from the Bayes rule:

$$P(I = 1 \mid \alpha, r, \Theta, \bar{D}_i) = \frac{P(I = 1, \bar{D}_i \mid \alpha, r, \Theta)}{P(\bar{D}_i \mid \alpha, r, \Theta)} \tag{10}$$

Equation (7) and the definition of likelihood lead to:

$$\begin{aligned} P(\bar{D}_i \mid \alpha, r, \Theta) &= a(\alpha, r, \bar{D}_i) \left(\prod_{j=1}^{x_i - 1} (1 - p_{i,j}) \right) (p_{i,x_i} * b(\alpha, r, \bar{D}_i) + (1 - p_{i,x_i})) \\ P(I = 1, \bar{D}_i \mid \alpha, r, \Theta) &= a(\alpha, r, \bar{D}_i) \left(\prod_{j=1}^{x_i - 1} (1 - p_{i,j}) \right) p_{i,x_i} b(\alpha, r, \bar{D}_i) \end{aligned} \tag{11}$$

Therefore, probability of churn for given customer after the end of observed period is:

$$P(I = 1 \mid \alpha, r, \Theta, \bar{D}_i) = \frac{p_{i,x_i} b(\alpha, r, \bar{D}_i)}{p_{i,x_i} * b(\alpha, r, \bar{D}_i) + (1 - p_{i,x_i})} \tag{12}$$

Parameter estimation via sequence of binary classifiers

In this section we describe the method of estimation of model's parameters. From equation (7) it follows that the model's likelihood depends on parameters α, r, Θ. Estimation of α, r is could be done by any known method of optimization. However,the estimation of Θ via direct application of optimization methods could be done by only under assumptions about parametric form of $p(dropout \mid \bar{D}_{i,j}, \Theta)$.

Here we propose different approach, that is free from such assumptions and allows to utilize non-parametric forms of $p(dropout \mid \bar{D}_{i,j}, \Theta)$. First, we use Minimization-Minimization method to derive special function, so called "surrogate function" such as it's extreme converges to log likelihood's extreme and then infer the solution of "surrogate" optimization problem as a sequence of binary classifiers. This way gives ability to plug in almost any method of binary classification as a solution for this sequence, therefore we aren't restricted to the particular form of $p(dropout \mid \bar{D}_{i,j}, \Theta)$.

Minimization-Minimization algorithm

In this section we briefly describe Minimization-Minimization method and the resulting algorithm. Details about the method could be found in (Hunter and Lange 2004) as well as in many other sources. The main idea is to build boundary function (so-called "surrogate function") which is less or equal than log likelihood and then, iteratively, build the sequence of parameters such as the maximum of surrogate function converges to the maximum of log likelihood. It turns out, that we can represent our particular optimization problem of our specific surrogate function as the weighted binary classification problem. Therefore, solution could be obtained via sequence of binary classifiers.

According to (Hunter and Lange 2004), we construct surrogate function $Q(\Theta, \tilde{\Theta}, \alpha, r)$ such as:

$$\begin{aligned} \log LL(\alpha, r, \Theta) &= Q(\Theta, \Theta, \alpha, r) \\ \log LL(\alpha, r, \Theta) &\geq Q(\Theta, \tilde{\Theta}, \alpha, r) \end{aligned} \tag{13}$$

for any Θ and $\tilde{\Theta}$.

Then, we consequently maximize function $Q(\Theta, \tilde{\Theta}, \alpha, r)$ over the first argument and substituting the result to the second argument until convergence. Then, parameters α and r are found over standard optimization procedure, then we repeat the whole procedure again until convergence:

Algorithm 1 MM algorithm

```
repeat
    repeat
        Θ := arg max_Θ Q(Θ, Θ̃, α, r)
        Θ̃ := Θ
    until Convergence
    (α, r) := arg max_{α,r} Q(Θ, Θ, α, r)
until Convergence
```

Surrogate function

In this section we derive surrogate function which is necessary for optimization via MM algorithm. We build surrogate function $Q_i(\Theta, \tilde{\Theta}, \alpha, r)$ on individual level, then obtain sample surrogate function as a sum:

$$Q(\Theta, \tilde{\Theta}, \alpha, r) = \sum_{i=1}^{N} Q_i(\Theta, \tilde{\Theta}, \alpha, r) \tag{14}$$

To construct customers surrogate function we apply Jensen inequality to equation (9):

$$\begin{aligned} Q_i(\Theta, \tilde{\Theta}, \alpha, r) &= S_i(\Theta, \tilde{\Theta}, \alpha, r) + \eta_i(\tilde{\Theta}, \alpha, r) \\ S_i(\Theta, \tilde{\Theta}, \alpha, r) &= \left(\sum_{j=1}^{x_i - 1} \log(1 - p_{i,j}) \right) + (1 - \mu_i(\tilde{\Theta}, \alpha, r)) \log(1 - p_{i,x_i}) + \mu_i(\tilde{\Theta}, \alpha, r) \log p_{i,x_i} \end{aligned} \tag{15}$$

where

$$\mu_i(\tilde{\Theta}, \alpha, r) = \frac{\tilde{p}_{i,x_i} b_i(\alpha, r)}{\tilde{p}_{i,x_i} b_i + (1 - \tilde{p}_{i,x_i})} \tag{16}$$

$$\begin{aligned} &\eta_i(\tilde{\Theta}, \alpha, r) = \log(\tilde{p}_{i,x_i} b_i(\alpha, r) + (1 - \tilde{p}_{i,x_i})) - \\ &- \left((1 - \mu_i(\tilde{\Theta}, \alpha, r)) \log(1 - \tilde{p}_{i,x_i}) + \mu_i(\tilde{\Theta}, \alpha, r) \log \tilde{p}_{i,x_i} \right) + \\ &+ \log a_i(\alpha, r) \end{aligned} \tag{17}$$

$$\begin{aligned} \tilde{p}_{i,j} &= p(dropout \mid \bar{D}_{i,j}, \tilde{\Theta}) \\ p_{i,j} &= p(dropout \mid \bar{D}_{i,j}, \Theta) \end{aligned} \tag{18}$$

Proof of that function satisfies conditions (13) could be found in Appendix B.

The term $S_i(\Theta, \tilde{\Theta}, \alpha, r)$ is the only component of $Q_i(\Theta, \tilde{\Theta}, \alpha, r)$ which depends on Θ, therefore optimization of $\sum_{i=1}^{N} Q_i(\Theta, \tilde{\Theta}, \alpha, r)$ by Θ is reduced to the optimization of $\sum_{i=1}^{N} S_i(\Theta, \tilde{\Theta}, \alpha, r)$.

Optimization of surrogate function via fitting binary classifier

In this section will reduce the problem of optimization of (15) by Θ to well-studied problem of fitting binary classifier. We will construct binary classification problem in a way that $-S_i(\Theta, \tilde{\Theta}, \alpha, r)$ from equation (15) is equal to weighted cross-entropy loss function. Therefore, the solution of this classification problem will deliver the solution of surrogate optimization problem.

Binary classification problem is:

$$\Theta = \arg\max_{\Theta} \sum_{i=1}^{N} \sum_{j=1}^{x_i} (w^1_{i,j} Y_{i,j} \log p(Y = 1 \mid \bar{D}_{i,j}, \Theta) + w^0_{i,j} (1 - Y_{i,j}) \log (1 - p(Y = 1 \mid \bar{D}_{i,j}, \Theta))) \tag{19}$$

where $Y_{i,j}$ are target variables, $w^1_{i,j}$ and $w^0_{i,j}$ are weights:

$$\begin{aligned} Y_{i,j} &= \begin{cases} 1, & \text{if } j = x_i \\ 0, & \text{if } 1 \leq j < x_i \end{cases} \\ w^1_{i,j} &= \begin{cases} \mu_i(\tilde{\Theta}, \alpha, r), & \text{if } j = x_i \\ 0, & \text{if } 1 \leq j < x_i \end{cases} \\ w^0_{i,j} &= \begin{cases} 1 - \mu_i(\tilde{\Theta}, \alpha, r), & \text{if } j = x_i \\ 1, & \text{if } 1 \leq j < x_i \end{cases} \end{aligned} \tag{20}$$

Substitution of (20) into (19) yields $\sum\limits_{i=1}^{N} S_i(\Theta, \tilde{\Theta}, \alpha, r)$, where $S_i(\Theta, \tilde{\Theta}, \alpha, r)$ is from (15). Therefore, optimizing $Q(\Theta, \tilde{\Theta}, \alpha, r)$ is the same as fitting binary classifier.

Any type of binary classifier $p(Y = 1 \mid \bar{D}_{i,j}, \Theta)$ with weighted cross-entropy loss could be plugged into (19) to get the solution of the problem. For example, it's possible to plug in logistic regression as well as tree-based methods, such as simple decision tree or ensemble of trees. In the next section we will demonstrate performance of the model then binary classifier is in the form of gradient-boosted trees.

Empirical validation

To validate performance of the model, we compared performance of churn prediction of original BG/NBD model and our model. We use transactional data from online retailer CDNOW. The dataset represents cohort of customers who made their first online purchase at CDNOW site from January to March 1997. The observed period is from 1997-01-01 to 1998-06-30. Dataset includes 69659 transactions of 23570 customers. Further details about the dataset could be found in (Fader and Hardie 2001). Before supplying dataset to the model, we aggregated purchases on the day/customer level since both models have minimum time frequency of 1 day. After aggregation the number of transactions reduced to 67591. To validate performance of both models, we compared ability to predict absence of transactions on validation period from the data in calibration period. We build 6 pairs of calibration-validation periods by splitting overall period into 2, as in Table 1.

	Calibration start	Calibration end	Validation end	Days
1	1997-01-01	1997-10-01	1998-06-30	273 / 272
2	1997-01-01	1997-11-12	1998-06-30	315 / 230
3	1997-01-01	1997-12-24	1998-06-30	357 / 188
4	1997-01-01	1998-02-04	1998-06-30	399 / 146
5	1997-01-01	1998-03-18	1998-06-30	441 / 104
6	1997-01-01	1998-04-29	1998-06-30	483 / 62

Table 1: Calibration/validation periods

On every pair we fit both models on calibration period only, calculated churn prediction by the end of calibration period and then measured area under ROC curve (AUC) metric of this prediction against the actual absence/presence of customer's transactions on validation period. The data from validation periods was not used during fitting on both models. This fact highlights distinction of the models from usual supervised methods. To fit GB/NDB model and calculate probability of churn (as $1 - p(alive)$) we used R package BTYD 2.4 (R version 3.3.2). To fit our model we used our implementation of (Algorithm 1). The binary classifier in step 3 in this algorithm was implemented via gradient boosting classifier with weighted cross-entropy loss function. For these purposes we used R package xgboost version 0.4-4. For binary classifier, we utilized time-dependent covariates for every transaction from the data available for the customer from the period prior to transaction. Covariates are provided in Table 2

Covariate	**Description**
bynow_avgd	Average period between transactions for given customer until the current tran
bynow_maxd	Maximum period between transactions for given customer until the current tra
bynow_mind	Minimum period between transactions for given customer until the current tra
bynow_avgsales	Average monetary value for given customer until the current transaction
bynow_minsales	Minimum monetary value for given customer until the current transaction
bynow_maxsales	Maximum monetary value for given customer until the current transaction
bynow_trans	Number of transactions for given customer until the current transaction
bynow_tenure	Days from the start of the period to the current transaction
dow	Transaction's day of week (from 1 to 7)
month	Transaction's month (from 1 to 12)
sales	Transaction's monetary value

Table 2: Covariates

To estimate confidence intervals for AUC we used percentile bootstrap intervals. The bootstrap procedure was performed as follows: For every pair of calibration/validation periods we draw 100 samples with repetition from the set of all customers and then performed model fitting and ACU calculation on the subset which belonged to these customers only.

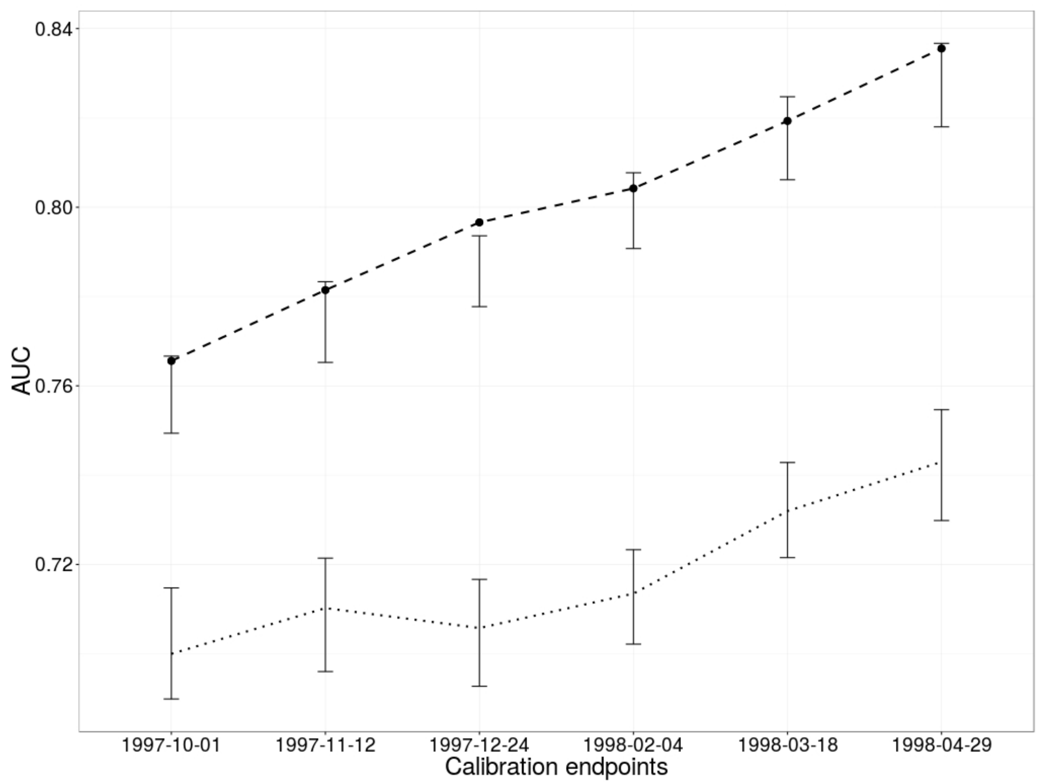

Figure 1: Bootstrap estimations of AUC. Dashed line represents our model, dotted line represents BG/NBD model, bars represent 95% percentile confidence intervals.

The numeric values are provided in the Table 3.

	Proposed model			BG/NBD model		
Pair	**95% CI**		**AUC**	**95% CI**		**AUC**
1	0.7493947	0.7666556	0.7655838	0.6898807	0.7147571	0.6999925
2	0.7652551	0.7833171	0.7814463	0.6960283	0.7214137	0.7102302
3	0.7777401	0.7935890	0.7966108	0.6927215	0.7166668	0.7057671
4	0.7907538	0.8077121	0.8042113	0.7021700	0.7233010	0.7134805
5	0.8061596	0.8247394	0.8193231	0.7215339	0.7428352	0.7319606
6	0.8180095	0.8366947	0.8355411	0.7298490	0.7546552	0.7429705

Table 3: Calibration/validation periods

Our model consistently outperforms BG/NBD model in terms of AUC metric for churn prediction on every pair. For comparison, Figure 2 shows ROC plot for both models for first calibration/validation pair (separation on 1997-10-01).

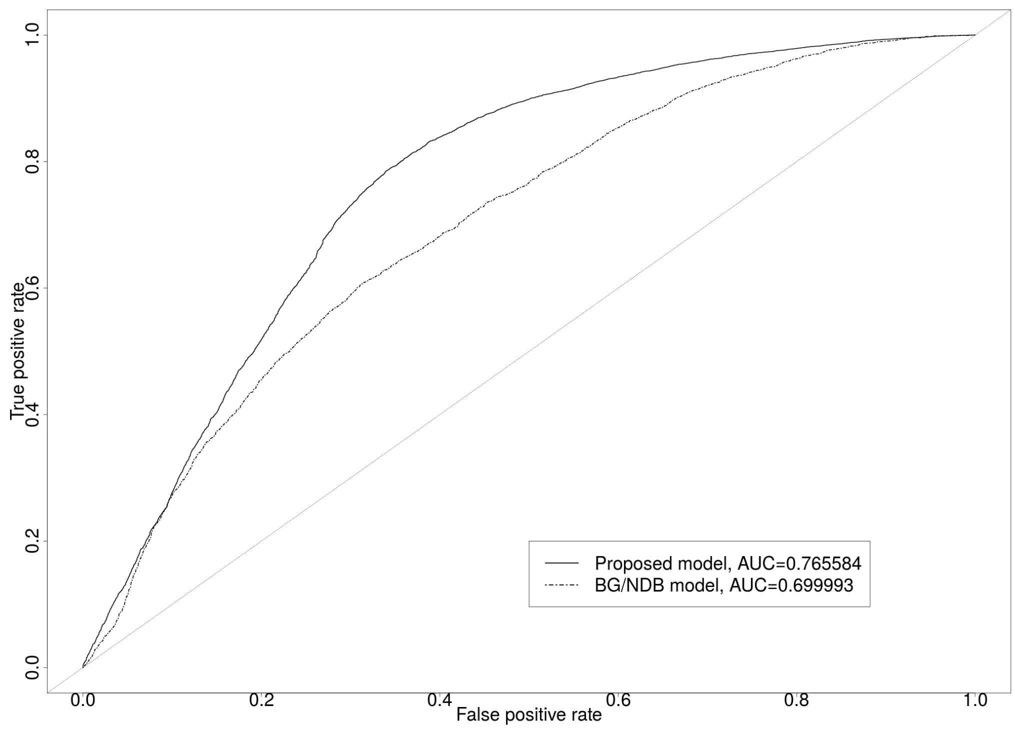

Figure 2: ROC curves. Solid line represents our model, dashed line represents BG/NBD model.

Figures 3 , 4 , 5 demonstrate the shape of log likelihood function where parameter Θ is fixed at the optimal value.

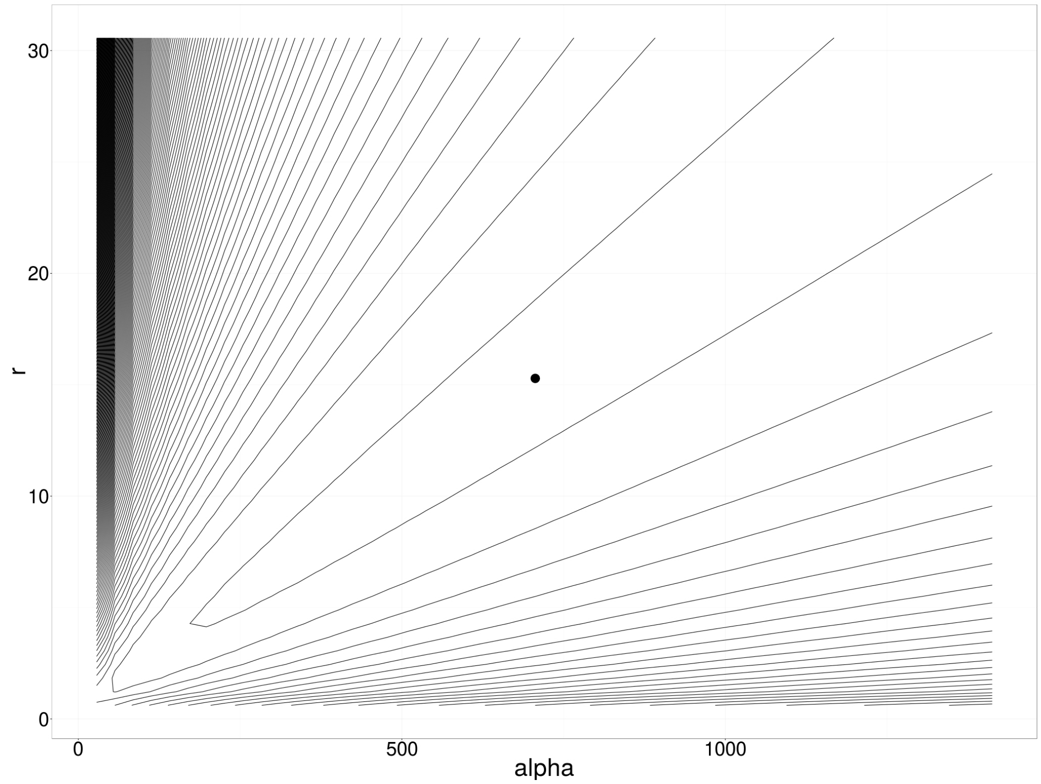

Figure 3: Level curves for the sample log likelihood. Horizontal axis corresponds to α, vertical to r. Black dot is at the optimal point.

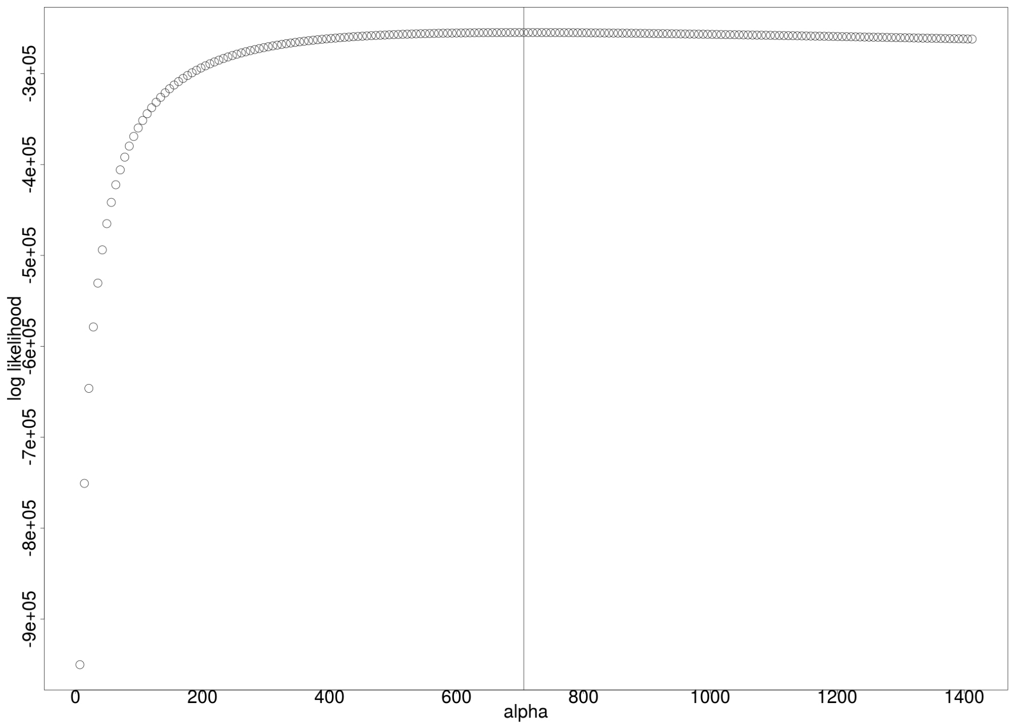

Figure 4: Log likelihood against α while r is fixed at optimal point.

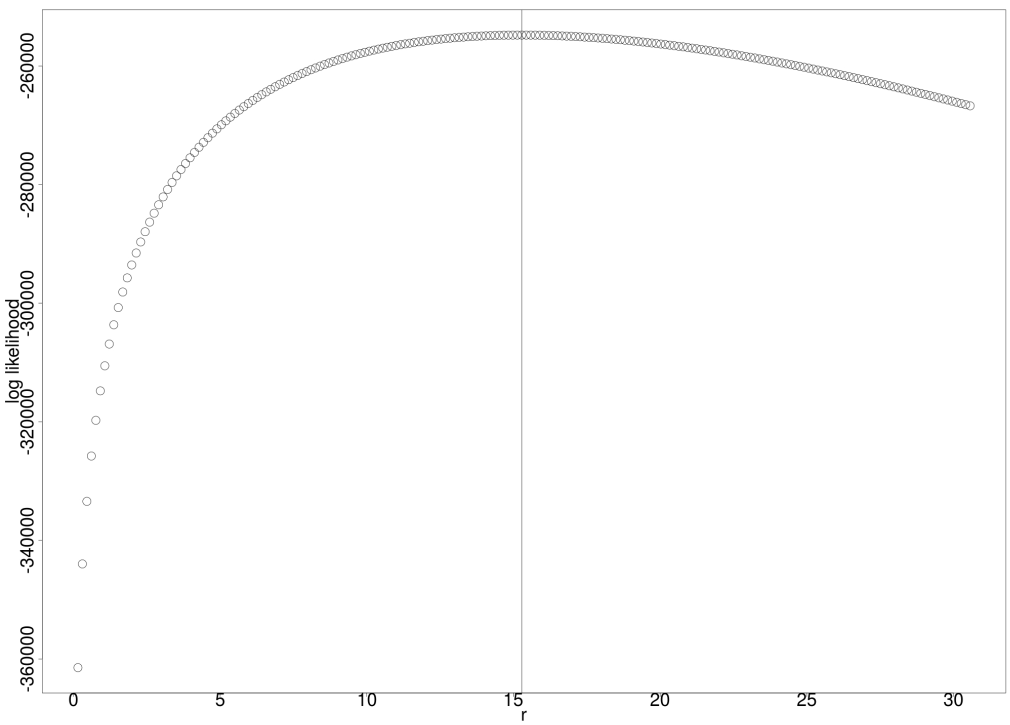

Figure 5: Log likelihood against r while α is fixed at optimal point.

Conclusion

In this article we presented an extension of BG/NBD model which allows to incorporate any kind of covariates, including time-dependent, as predictors for the probability of dropout. To do so, we employed novel approach which allows to reduce unsupervised problem to the converging sequence of supervised classification problems. To empirically estimate the predictive power of the model, we have compared churn prediction metrics from proposed and BG/NBD model and found that proposed model systematically outperforms original model.

Appendix A. Expectation of individual likelihood

Substitution of (5) into (6) yields:

$$L(\alpha, r, \Theta \mid \bar{D}_i) = \\ \frac{\alpha^r}{\Gamma(r)} \left(\prod_{j=1}^{x_i-1} (1 - p_{i,j}) \right) \int\limits_0^\infty \lambda^{x_i+r-1} e^{-\lambda(t_{i,x_i}+\alpha)} (p_{i,x_i} + (1 - p_{i,x_i}) e^{-\lambda c_i})\, d\lambda \tag{A.1}$$

This integral could be expressed via gamma function, therefore:

$$L(\alpha, r, \Theta \mid \bar{D}_i) =$$

$$= \frac{\alpha^r}{\Gamma(r)} \left(\prod_{j=1}^{x_i-1} (1 - p_{i,j}) \right) \left(p_{i,x_i} \frac{\Gamma(x_i + r)}{(t_{i,x_i} + \alpha)^{x_i+r}} + (1 - p_{i,x_i}) \frac{\Gamma(x_i + r)}{(t_{i,x_i} + c_i + \alpha)^{x_i+r}} \right) =$$

$$= \alpha^r \frac{\Gamma(x_i + r)}{\Gamma(r)} (t_{i,x_i} + c_i + \alpha)^{-(x_i+r)} \left(\prod_{j=1}^{x_i-1} (1 - p_{i,j}) \right) \left(p_{i,x_i} \big(1 + \frac{c_i}{t_{i,x_i} + \alpha}\big)^{x_i+r} + (1 - p_{i,x_i}) \right) \tag{A.2}$$

This this equation immediately yields to (7).

Appendix B. The proof of conditions (13)

Equation (15) yields:

$$Q_i(\Theta, \tilde{\Theta}, \alpha, r) = \\ = \left(\sum_{j=1}^{x_i-1} \log(1 - p_{i,j}) \right) + (1 - \mu_i(\tilde{\Theta}, \alpha, r)) \log(1 - p_{i,x_i}) + \mu_i(\tilde{\Theta}, \alpha, r) \log p_{i,x_i} + \eta_i(\tilde{\Theta}, \alpha, r) = \\ = \log a_i(\alpha, r) + \left(\sum_{j=1}^{x_i-1} \log(1 - p_{i,j}) \right) + \log(\tilde{p}_{i,x_i} b_i(\alpha, r) + (1 - \tilde{p}_{i,x_i})) + \\ + \frac{\tilde{p}_{i,x_i} b_i(\alpha, r)}{\tilde{p}_{i,x_i} b_i(\alpha, r) + (1 - \tilde{p}_{i,x_i})} \log \frac{p_{i,x_i}}{\tilde{p}_{i,x_i}} + \frac{1 - \tilde{p}_{i,x_i}}{\tilde{p}_{i,x_i} b_i(\alpha, r) + (1 - \tilde{p}_{i,x_i})} \log \frac{1 - p_{i,x_i}}{1 - \tilde{p}_{i,x_i}} \tag{B.1}$$

Applying Jensen inequality to the last 2 terms yields:

$$\begin{aligned}
&\frac{\tilde{p}_{i,x_i} b_i(\alpha,r)}{\tilde{p}_{i,x_i} b_i(\alpha,r) + (1-\tilde{p}_{i,x_i})} \log \frac{p_{i,x_i}}{\tilde{p}_{i,x_i}} + \frac{1-\tilde{p}_{i,x_i}}{\tilde{p}_{i,x_i} b_i(\alpha,r) + (1-\tilde{p}_{i,x_i})} \log \frac{1-p_{i,x_i}}{1-\tilde{p}_{i,x_i}} \leq \\
&\leq \log\left(\frac{\tilde{p}_{i,x_i} b_i(\alpha,r)}{\tilde{p}_{i,x_i} b_i(\alpha,r) + (1-\tilde{p}_{i,x_i})} \frac{p_{i,x_i}}{\tilde{p}_{i,x_i}} + \frac{1-\tilde{p}_{i,x_i}}{\tilde{p}_{i,x_i} b_i(\alpha,r) + (1-\tilde{p}_{i,x_i})} \frac{1-p_{i,x_i}}{1-\tilde{p}_{i,x_i}}\right) = \\
&= \log\left(p_{i,x_i} b_i(\alpha,r) + (1-p_{i,x_i})\right) - \log\left(\tilde{p}_{i,x_i} b_i(\alpha,r) + (1-\tilde{p}_{i,x_i})\right)
\end{aligned} \tag{B.2}$$

Substitution of (B.2) into (B.1) immediately yields to (13)

References

[1] Koen W. De Bock and Dirk Van den Poel. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, 38(10):12293 – 12301, 2011. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2011.04.007. URL http://www.sciencedirect.com/science/article/pii/S0957417411005239.

[2] Kristof Coussement and Dirk Van den Poel. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3, Part 2):6127–6134, 2009. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2008.07.021. URL http://www.sciencedirect.com/science/article/pii/S095741740800479X.

[3] Richard Colombo David C. Schmittlein, Donald G. Morrison. Counting your customers: Who are they and what will they do next? *Management Science*, 33(1):1–24, 1987. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2631608.

[4] David C. Schmittlein Donald G. Morrison. Generalizing the nbd model for customer purchases: What are the implications and is it worth the effort. *Journal of Business & Economic Statistics*, 6(2):145–159, 1988. ISSN 07350015. URL http://www.jstor.org/stable/1391551.

[5] A. S. C. Ehrenberg. The pattern of consumer purchases. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 8(1):26–41, 1959. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2985810.

[6] Peter S. Fader and Bruce G. S. Hardie. Forecasting repeat sales at cdnow: A case study. *Interfaces*, 31(3_supplement):S94–S107, 2001. doi: 10.1287/inte.31.3s.94.9683. URL http://dx.doi.org/10.1287/inte.31.3s.94.9683.

[7] Peter S. Fader and Bruce G. S. Hardie. Incorporating time-invariant covariates into the pareto/nbd and bg/nbd models. *Working paper*, 2007. URL http://www.brucehardie.com/notes/019/.

[8] Peter S. Fader and Bruce G. S. Hardie. Computing p(alive) using the bg/nbd model. *Working paper*, 2008. URL http://www.brucehardie.com/notes/021/.

[9] Peter S. Fader and Bruce G.S. Hardie. Probability models for customer-base analysis. *Journal of Interactive Marketing*, 23(1):61 – 69, 2009. ISSN 1094-9968. doi: http://dx.doi.org/10.1016/j.intmar.2008.11.003. URL http://www.sciencedirect.com/science/article/pii/S1094996808000108. Anniversary Issue.

[10] Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. counting your customers the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 24(2):275–284, 2005. doi: 10.1287/mksc.1040.0098. URL `http://dx.doi.org/10.1287/mksc.1040.0098`.

[11] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004. doi: 10.1198/0003130042836. URL `http://dx.doi.org/10.1198/0003130042836`.

[12] Ali Tamaddoni Jahromi, Mohammad Mehdi Sepehri, Babak Teimourpour, and Sarvenaz Choobdar. Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *Journal of Strategic Marketing*, 18(7):587–598, 2010. doi: 10.1080/0965254X.2010.529158. URL `http://dx.doi.org/10.1080/0965254X.2010.529158`.

[13] Ali Tamaddoni, Stanislav Stakhovych, and Michael Ewing. Comparing churn prediction techniques and assessing their performance a contingent perspective. *Journal of Service Research*, 19(2):123–141, 2016.

[14] Xiaobing Yu, Shunsheng Guo, Jun Guo, and Xiaorong Huang. An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, 38(3):1425 – 1430, 2011. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2010.07.049. URL `http://www.sciencedirect.com/science/article/pii/S0957417410006779`.