Churn Analysis

Josep Curto Díaz, Adjunct Professor[a]

[a]IE Business School, Madrid, 28006, Spain

This version was compiled on March 28, 2020

This technical note introduces what is Churn Analysis and its benefits and limitations.

churn analysis | survival analysis | customer analytics | r

The problem

- What is the worst that could happen to a company? **Losing customers**!
- **Goal**: understand how our customers behave and define actions based on their behaviour. In particular, any behaviour with a hint of attrition.

Definition

- **Churn**, or attrition, refers to an existing customer deciding to end the business relationship.
- Churn analysis aims to divide customers in *active*, *inactive* and *about to churn*.
- Churn models predict probability of churn given influencing factors or key factors.
- If action is taken to address the factors that influence churn, the model in turn becomes obsolete and must be rebuilt with new churn data and influencing factors (this is what we call **concept drift**, a change in the relationships between input and output data in the underlying problem over time).

Main Concepts in Churn

There are two types of churn:

- **Voluntary churn**: occurs due to a decision by the customer to switch to another company or service provider.
- **Involuntary churn**: occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location.

Involuntary reasons for churn must be excluded from the analytical models as most of the times can not be influenced.

When analysis churn all customer data is interesting:

- CRM data
- Device usage
- Interaction data
- Clickstream data
- Granular credit card
- Insurance data
- ...

Companies focus on **identifying**, **understanding**, **predicting**, **reducing** and/or **managing** churn.

How to apply churn analysis

- **[BU]** Determine business needs: churn identification, churn understanding, churn prediction and/or churn reduction.
- **[DU]** Data Sourcing, Cleaning & Exploration
- **[M]** Select specific technique(s)
- **[M]** Applied specific technique(s)
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Reasons for churn

The hardest part is to identify reasons for churn. Many companies use NPS (Net Promoter Score) as starting point. NPS is not enough. We require to determine the importance and weight of the detected factors. This can be done using Structural Modelling Equations (SME) and other predictive modelling techniques.

What is SEM (Structural Equation Modelling)? Read the following article.

Learning and understing SEM is beyond the scope of this course. In case you are interested, you can consider the following R packages for SEM:

- Lavaan (Latent Variable Analysis)
- SEM (Structural Equation Models)
- ctsem (Continuous Time Structural Equation Modelling)

Benefits

- Determine if churn is a real problem in the organization
- Understand the churn causes
- Analyze and establish churn profiles and identify critical moments
- Combine CLV and churn analysis for profiling
- Awake inactive customers based on specific marketing actions
- Retain customers based on specific marketing actions
- Invite customers to churn based on specific marketing actions (or lack of customer engagement)
- Improve customer experience
- Predict future customer behaviour

Use Cases

Some examples:

- **Telecom Providers**: anticipating subscription cancellations and proposing specific commercial actions to foster loyalty
- **E-commerce Players**: increasing loyalty and client lifetime value by activating personnalized campaigns to "dormant" clients - pushing the right product at the right time through the right channel
- **Banking & Insurance Companies**: predicting life events from behavioral data to anticipate structural changes in the client's consumption profile that may signal churn or upsell/cross-sell opportunities
- **Generic**: create alarms (based on churn changes), create churn critical path (when, where, how),...

About predicting churn

We already know that we have several escenarios:

- **Contractual**: Customers make purchases at discrete intervals, on a contract or autoplay. Cancellation event is observed and recorded.
- **Non-contractual**: Customers are free to buy or not at any time. Churn event is not explicitly observed.
- **Voluntary**: Customers make the choice to leave the service.

- **Involuntary**: Customers ara forced to discontinue service and/or payments.

  Note: Involuntary/voluntary churn can be present in either contractual or non-contractual settings.

A **predictive churn model** is a measure of the immediate or future risk of a customer cancellation. The goal of the model is reduce churn.

There are many different techniques for predicting churn. The following article is recommended: Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models.

In general, classification algorithms can be used to predict the churn status. For instance, we can use random forest (an ensemble of decision trees) for churn prediction.

About Survival Analysis

Survival Analysis is a technique that can be used to understand churn evolution. It is know as well as **failure time analysis**. It is the analysis of time to death. It can be used anywhere you want to know what factors affect the time for an event to occur:

- Churn
- Germination timing
- Arrival of a migrant or parasite
- Dispersal of seeds or offspring
- Failure time in mechanical systems
- Response to stimulus

When working with survival analysis, we can find ourselves dealing with missing data. This is called in this context: **censoring**. We have two situations:

- **Right censoring**: Where the date of death is unknown but is after some known date. For example,
  - Date of death is after the end of the study
  - Subject is removed from the study (patient withdraws, animal escapes, plant gets eaten, etc.)
- **Left censoring**: Occurs when a subject's survival time is incomplete on the left side of the follow-up period.
  - Following up a patient after being tested for an infection, we don't know the exact time of exposure

Survival analysis consist in defining and analyzing **The Survival function**. The Survival function S is the probability that the time of death T is greater than some specified time t:

$$S(t) = Pr(T > t) \quad [1]$$

It is composed of:

- **The underlying Hazard function**: How the risk of death per unit time changes over time at baseline covariates.
- **The effect parameters**: How the hazard varies in response to the covariates.

The **Kaplan–Meier estimator** is one of the most frequently used methods of survival analysis. The estimate may be useful to examine recovery rates, the probability of death, and the effectiveness of treatment. It is limited in its ability to estimate survival adjusted for covariates; parametric survival models and the Cox Proportional hazards model may be useful to estimate covariate-adjusted survival.

The Kaplan–Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common estimators is Greenwood's formula:

$$\widehat{\mathrm{Var}}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{i:\ t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \quad [2]$$

where d_i is the number of cases and n_i is the total number of observations, for $t_i < t$.

In some cases, one may wish to compare different Kaplan–Meier curves. This can be done by the log rank test, and the Cox proportional hazards test. Other statistics that may be of use with this estimator are the Hall-Wellner band and the equal-precision band.

The Cox Proportional-Hazards Model is the most common model used to determine the effects of covariates on survival

$$h_i(t) = h_0(t) exp(\beta_1 x_{i1} + \beta_2 x_{ik} + ... + \beta_2 x_{ik}) \quad [3]$$

It is a semi-parametric model:

- The baseline hazard function is unspecified
- The effects of the covariates are multiplicative
- Doesn't make arbitrary assumptions about the shape/form of the baseline hazard function

The Proportional Hazards Assumption

- Covariates multiply the hazard by some constant
  - e.g. a drug may halve a subjects risk of death at any time t
- The effect is the same at any time point

Related Packages

- Survival Analysis Package
- plsRcox: Partial Least Squares Regression for Cox Models and Related Techniques
- Caret Package
- NestedCohort: Survival Analysis for Cohorts with Missing Covariate Information
- DDPGPSurv: DDP-GP Survival Analysis
- tranSurv: Estimating a Survival Distribution in the Presence of Dependent Left Truncation and Right Censoring
- survsup: beautiful survival curves
- survxai: Visualization of the Local and Global Survival Model Explanations
- did: Treatment Effects with Multiple Periods and Groups

References

- Churn Analysis
- An Introduction to Survival Analysis
- Zero Defections: Quality comes to service, Reichheld & Sasser. Harvard Business review. 1990.
- Van Den Poel; Lariviére (2004). Customer Attrition Analysis For Financial Services Using Proportional Hazard Models. European Journal of Operational Research 157: 196–217.
- Applying and evaluating models to predict customer attrition using data mining techniques, Tom Au, et al. Journal of Comparative International Management. 1 June 2003
- Mittal, Vikas and Sarkees, Matthew, Customer Divestment (2006). Journal of Relationship Marketing, 5(2/3), 71-85, 2006.
- Mittal, Vikas and Sarkees, Matthew and Murshed, Feisal, The Right Way to Manage Unprofitable Customers (April 1, 2008). Harvard Business Review, Vol. 86, No. 4, 2008.
- Buckinx Wouter, Dirk Van den Poel (2005), Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting, European Journal of Operational Research, 164 (1), 252-268.

- Tsymbal, A., The problem of concept drift: Definitions and related work. Technical Report. 2004, Department of Computer Science, Trinity College: Dublin, Ireland.