Paper 1131-2017

Application of Survival Analysis for Predicting Customer Churn with Recency, Frequency, and Monetary

Bo Zhang, IBM; Liwei Wang, Pharmaceutical Product Development Inc.

ABSTRACT

Customer churn is an important area of concern that affects not just the growth of your company, but also the profit. Conventional survival analysis can provide a customer's likelihood to churn in the near term, but it does not take into account the lifetime value of the higher-risk churn customers you are trying to retain. Not all customers are equally important to your company. Recency, frequency, and monetary (RFM) analysis can help companies identify customers that are most important and most likely to respond to a retention offer. In this paper, we use the IML and PHREG procedures to combine the RFM analysis and survival analysis in order to determine the optimal number of higher-risk and higher-value customers to retain.

INTRODUCTION

Customer importance can be quantified by measuring RFM:

- Recency - how recently did your customers make a purchase.
- Frequency - how often do your customers make a purchase.
- Monetary - how much have your customers made a purchase.

Based on such RFM analysis, a more recent, higher frequency and larger total monetary contribution customer is a higher-value customer. However, a higher-value customer of today may not be your star customer in the near future, or even not your customer anymore. This reveals the key drawback of the RFM analysis: it's only a descriptive analysis, and motivates us to propose a multilevel model shown in Figure 1.

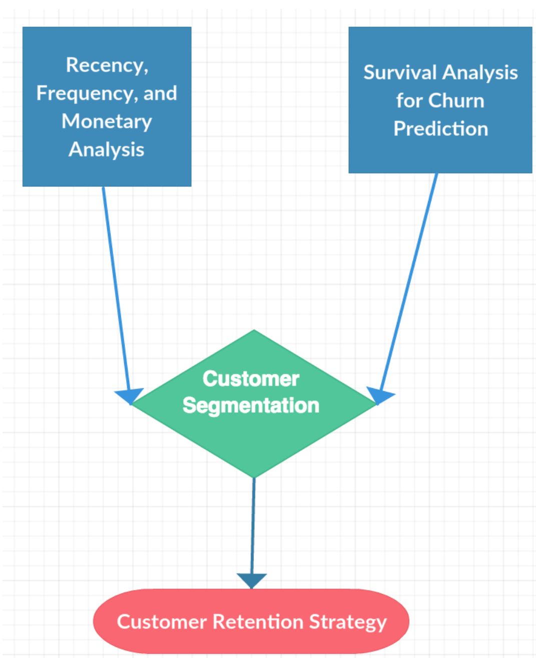

Figure 1. Multilevel Model Diagram

The first step, RFM analysis, provides hindsight and answers what has happened to your customers, but it's not able to create insight and answer what could happen. The second step, customer churn analysis, which is a predictive analysis of calculating the probability of churn for each customer, provides additional insights onto customer importance. It's critical to identify your higher-value but also higher-risk customers. Also, management can target different retention strategies for different customer segments (high/low recency, high/low frequency, high/low monetary, and high/low retention). Table 1 lists out sample columns of the example data used in this paper: ID, Tenure, Churn, Type, Country, Region, Industry, Marketing, Usage, Calls, ProductPurchased, ProductSeen, PurchaseDate, Amount, and so on.

Column	Description
ID	Customer ID
Tenure	How long the customer has been staying until now or until the customer churns
Churn	Whether the customer churned (with a one) or still staying (with a zero)
Type	Account Type
Country	Customer Country
Region	Customer Region
Industry	Customer Industry
Marketing	Marketing Campaign
Usage	Customer Usage
Calls	Customer Calls
ProductPurchased	Number of Products Purchased
ProductSeen	Number of Products Seen
PurchaseDate	Purchase Dates
Amount	Sales Amount
…	…

Table 1. Sample Data Columns

RECENCY, FREQUENCY, MONETARY ANALYSIS

A SAS macro function is created to calculate three KPIs: recency, frequency and monetary score. But for each KPI, instead of returning quintile ranking of each observation, the function calculates a normal curve equivalent (NCE) score. NCE technique fits a normal distribution (preserving the equal-interval properties of a z-score) of the data for each KPI into a 0-100 scale (similar to percentile-rank). To put together with the survival analysis later, we don't want to simply split the observations into quintiles and aggregate them together. We want to standardize the KPIs in a way that later can be legitimately averaged with retention scores. The formula to calculate NCE is 21.063 * z-score +50. Table 2 shows the NCE conversion for some observations based on monetary as an example.

Monetary	Percentile Rank	NCE_M
1413	1	1
1517	12	25
1584	31	40
1648	50	50
1812	86	73
1912	99	99

Table 2. Normal Curve Equivalent (NCE) Conversion

The arguments expected, options available and outputs returned in the macro function are summarized in Table 3.

Object	Type (Values)
inputData	Argument
ID	Argument
PurchaseDate	Argument
Amount	Argument
Standardization	Option (Quintile, PercentileRank, NCE)
outputData	output

Table 3. Macro Function Arguments, Options and Outputs

SURVIVAL ANALYSIS FOR CHURN PREDICTION

As part of the efforts to design retention strategy for different customer segments, we model the "time to churn" in order to determine the factors associated with customers who churned. Survival analysis is commonly adopted when the target is to predict when certain event will happen. While it’s widely used in clinical trials study, it can easily be transformed in a way that can be used in customer churn.

Figure 2 shows sample rows and columns of the data, including how long the customer has been around until now or until the customer churns, whether the customer churned (with 1) or not (with 0), Account Type, Country, Region, Industry, Marketing, and so on.

Tenure	Churn	Type	Country	Region	Industry	…	Marketing	Usage	Calls	…
13	1	1	14	12	07	…	3	10	28	45
11	1	1	13	06	02	…	1	13	15	24
68	0	1	02	07	02	…	4	8	10	23
33	1	0	13	09	01	…	1	2	3	3
23	1	0	10	19	03	…	2	1	1	1
41	0	1	19	25	04	…	1	9	17	9
…	…	…	…	…	…	…	…	…	…	…

Figure 2. Sample Rows and Columns of the Data

Let’s take a look at four customers who started about the same time as an example in Figure 3. Customer A churned after about 3 months and hence we are able to measure the "time to churn" for this particular customer. Customer B churned after about 6 months so the "time to churn" can be estimated as well. However, for customers C and D, since they are still being customers, we don’t know what’s the time to churn for them. Hence, we need to rely on the survival model adapted for such censored situations. The survival model can identify what's driving the probability to churn, even for the customers who didn't churn yet. Because the fact that they are still in business today still conveys some information.

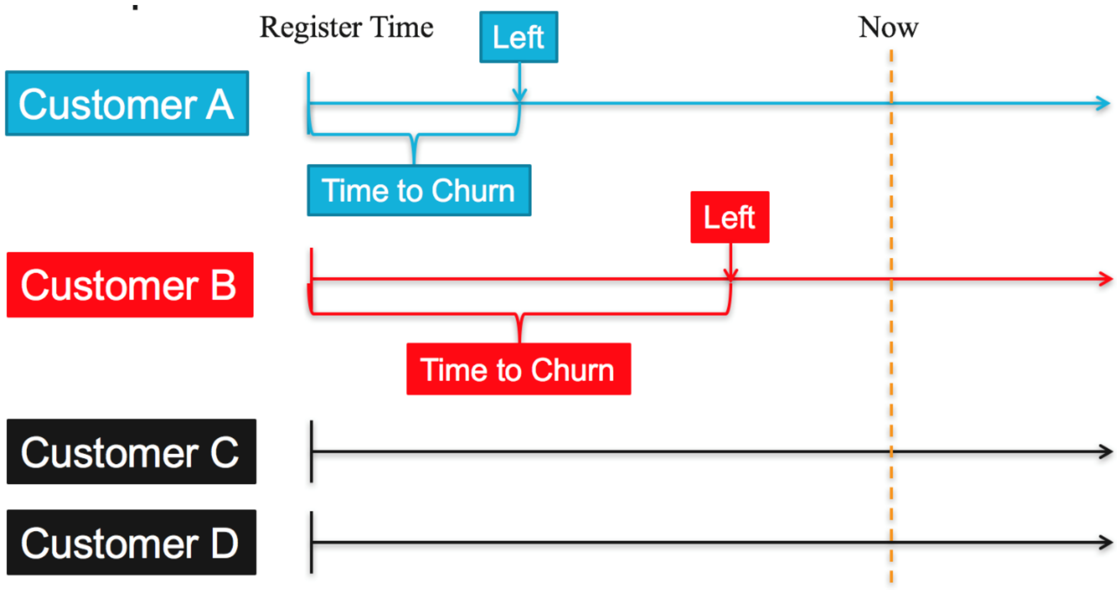

Figure 3. Four Example Customers

Now let's review the steps of this survival analysis pipeline, which is summarized in Figure 4.The first step of this pipeline of course involves obtaining our raw data in Table 1. These columns can come from a variety of data sources, including marketing data, user behavior data, social media data, survey data and so on. Then we need to extract useful columns from all of them. Feature extraction allows us to incorporate domain knowledge on what's driving the churn probability when representing each of the customers. Once a good representation for the customers is obtained, we're able to perform survival analysis. And this involves training a survival model using PROC PHREG with different options. Once we determined a model, we then evaluate how well it's performing based on the lift measurement. If the model is not performing well, we can extract new features and use different options in PROC PHREG. Finally, when a final model is determined, we can implement it to make predictions on all customers. Eventually, all the customers have their retention NCE score by converting the 1-Churn score into NCE.

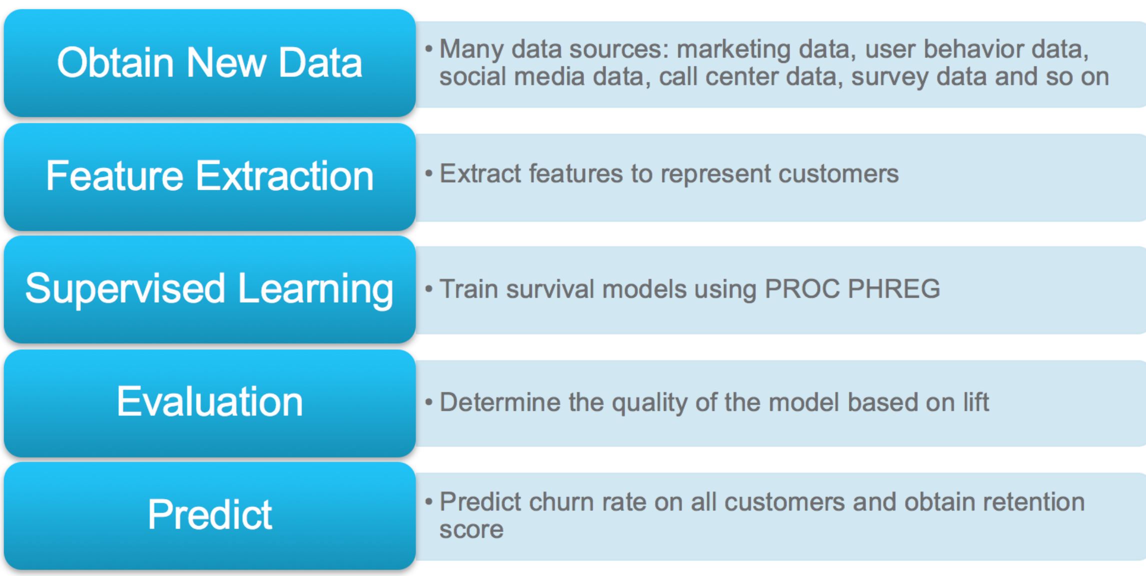

Figure 4. Survival Analysis Pipeline

CUSTOMER SEGMENTATION

Now you can perform customer segmentation based on the outputs from RFM analysis and survival analysis. Table 4 shows the Recency, Frequency, Monetary, Retention NCE Scores (RFMR) for some observations.

ID	NCE_Recency	NCE_Frequency	NCE_Monetary	NCE_Retention
A	4	2	6	3
B	19	65	15	78
C	60	34	35	42
D	90	17	18	15
E	43	89	76	89
F	98	98	96	25
…	…	…	…	…

Table 4. Recency, Frequency, Monetary, Retention NCE Scores

Based on this data, you can perform advanced clustering techniques to segment your customers, such as K-means and Hierarchical Clustering. Since all the four NCE scores can be legitimately averaged, you can even compare 1 score with all the rest scores, or 2 scores with the other two scores. Here in Table 5, we present several typical segments to illustrate how management can design different retention strategies for different segments.

Segmentation	Description	Retention Strategy
Best Loyal Customers	Purchased most recently, purchased most often, purchased the most, and had low likelihood to churn	No discount or coupon, upselling and cross-selling
Almost Lost Best Customers	Purchased most recently, purchased most often, purchased the most, but had high likelihood to churn	Aggressive pricing strategies
Almost Lost Cheap Customers	Last purchased quite a while, purchased very few, and purchased very little, and had high likelihood to churn	Don’t spend too much effort to retain
…	…	…

Table 5. Recency, Frequency, Monetary, Retention Segments

CONCLUSION

Recency, Frequency, Monetary analysis is a common prescriptive analysis technique which can help better understand your customer value. Integrating the RFM NCE scores with the retention NCE scores, obtained from the survival analysis, will help determine the optimal retention strategy for different segments with different risks and different values. We recommend that the RFMR segmentation is presented to the management team in a real-time fashion.

REFERENCES

Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. SAS User Group International Online Proceedings.

Lu, J., & Park, O. (2003). Modeling customer lifetime value using survival analysis—an application in the telecommunications industry. Data Mining Techniques, 120-128.

SAS/STAT(R) User's Guide.

Sung, C., Zhang, B., Higgins, C. Y., & Choe, Y. (2016). Data-Driven Sales Leads Prediction for Everything-as-a-Service in the Cloud. In Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on (pp. 557-563). IEEE.

Sung, C., Higgins, C. Y., Zhang, B., & Choe, Y. (2017). Evaluating Deep Learning in Churn Prediction for Everything-as-a-Service in the Cloud. International Joint Conference on Neural Networks (IJCNN).

Zhang, R., Liu, F. (2013). Explore RFM Analysis using SAS® Data Mining Procedures. SAS User Group International Online Proceedings.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bo Zhang
IBM
Email: bozhang@us.ibm.com