Implementing a Customer Life Time Value Framework in SAS®

Krishna Mehta, Thinkalytics

Abstract

In this paper we focus on the analytics of implementing CLTV in SAS®. We describe the different analytics steps needed to implement the CLTV framework, appropriate SAS procedures and some pitfalls to avoid. We first layout the data requirements and analytics framework for putting together a CLTV measure, and then discuss the different components in greater details. Two key components of the measure are (a) an estimation of the life time of the customer and (b) their projected revenue for each of the time periods over their lifetime. These are combined to then obtain the lifetime value.

For lifetime estimation, we compare the relative merits and drawbacks of using a Discrete Time Logistic Model (with PROC LOGISTIC), a Proportional Hazard Model (using PROC PHREG) and a Continuous Time Parametric Model (using PROC LIFEREG). We then present situations where one of the methods might be more appropriate.

For estimating future value, we discuss different approaches to revenue estimation, as well as cross sell and upsell models. We compare situations where it is appropriate to use simple logistic models (PROC LOGISTIC), to where more complex models that involve multinomial or generalized linear models (PROC GENMOD) are needed. We finally suggest circumstances where the modeler should consider a net return or Uplift model, and provide tips on implementing it in SAS using PROC LOGISTIC.

Introduction

As marketing maturity has grown over time, corporations have realized the value of taking a 360 degree view of their customers. Traditionally, acquisition, cross-sell/upsell and retention have been handled by different groups within a firm. These groups operate largely independently of one another and the end outcome is often sub-optimal. The organization stands to gain by focusing on prospects that will stay on longer and avoid the ones who are fickle or risky to serve. Similarly, it is better to focus the retention efforts and incentives on customers who are likely to buy more in the future. Therefore, organizations need to understand the life time value of their customers (CLTV).

In this paper, we first present an overview of the data and analytics framework required to setup a CLTV framework. Then we focus on the analytics and present alternatives and best practices for implementing two key components in a SAS platform (a) The estimation of the life time of a customer (b) The various approaches to estimating value for a given lifetime

Data Requirements and Analytics Framework

In this section we will discuss some of the data requirements for setting up the CLTV framework as well as the associated analytics. When we develop a measure for the life time value of a customer, we are focusing on two key elements (a) What is the best estimate of the time the customer is likely to be with the firm? (b) How much will we earn from the customer over this time period?
To answer the first question we need data related to subscription terms (in industries where relevant), customers usage behavior/shopping patterns, billing and payment information as well as customer contact/customer service issues. In addition, we need information on past customer attrition. We will use all of this information to develop a model that allows us to predict the survival/retention probability of the customer for future time periods.

To answer the second question we will need usage/shopping patterns, as well as customer specific information on demographics/firmographics, past cross sells and revenue streams. We will use this to develop propensity models for cross sell/upsell and estimated revenue models.

So to develop a CLTV framework, we need to start by understanding the different data elements that are available for the analysis. Here is a list of data that are generally needed for this type of exercise.

- Demographics
- Transactions/ Usage
- Payment
- Customer Service
- Attrition
- Cost

One of the biggest challenges to implementing a CLTV framework is to ensure that these data elements are available for analysis and also available in the environment of intended use. A good strategy often is to recognize the data limitations and design a limited CLTV implementation which gets enhanced as more data become available.

Let us next discuss the analytics framework that is needed for putting together a CLTV measure. As we mentioned above, one of the biggest challenges in the process is that we have to collect a lot of data from very different sources. Therefore, it is very important to evaluate the data quality carefully. In our experience, putting together the data from different sources can be a big challenge.

Next, we need to develop strategies for combining the different data elements and treating the data for anomalies like outliers and missing values. This is not very different from what would be done in any other modeling exercise and is not covered here. Interested readers can check our past NESUG papers on implementing appropriate missing imputation and outlier treatment strategies in SAS.

Once we get past the data preparation stage we need to consider the modeling that will be required for the process. Usually three types of models/algorithms need to be developed

(a) Develop models to understand how long the customer will remain a customer (some form of survival modeling)
(b) Develop models to understand the purchase behavior of the customer. These would include modeling for cross sell/upsell as well as modeling for revenue projection
(c) Develop schema/algorithms to segment the customers. This might be needed because we might find that the predictors that drive the retention or purchase behavior are significantly different across groups and cannot be captured by a single model.

In the following section, we will focus on modeling techniques and best practices around estimating the life time of the customer and the customer’s purchase behavior.
Before concluding this section, we would like to emphasize that in choosing between the different approaches to modeling and data preparation, special attention should be paid to implementability and automation friendliness. This is because a CLTV measure ends up being used by different people in the organization apart from the analytics groups, including marketing, strategists and even customer service personnel.

**Estimation of The Customers Lifetime**

In this section, we will discuss the alternatives for estimating the lifetime of a customer. In general modelers in this situation consider among three alternatives:

- Continuous Time Parametric Model (Using PROC LIFEREG)
- Proportional Hazard Model (Using PROC PHREG)
- Discrete Time Logistic Model (Using PROC LOGISTIC)

Continuous Time Parametric Models assume that survival probability follows certain distributional patterns and estimates the parameters of the distribution from the data. Future survival probabilities are then estimated using this model. This approach is very appealing because of the ease with which it can be implemented, but has a serious shortcoming in that it cannot accommodate predictors that change over time. This is a problem because more often than not, retention probability is directly influenced by events like change in employment status or some external economic shock, things that change over time. Another less critical shortcoming of this approach is that it is bound by the parametric constraints of the empirical distribution, and true data might now follow it.

A very popular technique to overcome this shortcoming is to develop a Proportional Hazard Model which uses Cox’s partial likelihood regression. The proportional hazard model is the most general of the regression models because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The model assumes that the underlying hazard rate (rather than survival time) is a function of the covariates. No assumption is made about the nature of the shape of the hazard function and the only assumption made is that hazard rates of the different individuals are proportional. While very appealing due to the reasons highlighted above, the proportional hazard model has some shortcomings. It takes significantly more computing resources to estimate and also is harder to validate due to the fact that PROC PHREG only provides estimates for covariates that do not vary over time.

Due to the above listed challenges of the Proportional hazard Model, a good compromise is to develop a discrete time logistic model using PROC LOGISTIC. It is constrained by the fact that it assumes a logistic distribution, but gives the user the flexibility of incorporating time varying independent variables. Another reason for going with this approach is that modelers are usually fairly familiar with PROC LOGISTIC, and have a lot of utilities developed around it, which they can continue to use.
Whether one uses PROC PHREG or PROC LOGISTIC one has to expand the dataset from having one observation per customer to one observation per customer per time period. This then allows us to include changes in covariates which vary over time.

Another thing to keep in mind for both these approaches is that while it is great to include time varying covariates, they do pose a serious challenge in implementation. If we are predicting customers retention probability over the next 24 months and it includes predictors that change over time, then we need to find a way to predict the future value of these predictors. This leads to the complexity of generating forecasts of these predictors or incorrectly assuming constant values for them.

One solution we have implemented which reduces this problem is to develop a family if discrete time logistic models with different time lags for the same covariate. For example, suppose we find that income helps predict the retention of a customer and we are interested in developing the retention probability over 24 months. Then we re-estimate the equation with income appearing in it with lag 1 to lag 24. And then use the equation where income appears with lag1 to predict retention probability 1 month in to future and lag24 to predict retention probability 24 months into the future.

To conclude the section, we would like to emphasize that each approach discussed in this section has its own set of limitations and a good understanding of all the constraints in model development and the implementation environment is required before a solution is finalized.
Estimating the Value of the Customer

Using the models in the previous section we can determine what the likely lifetime of our customer is, or more specifically what is the survival probability of a customer for different time periods. The next task is to get a value estimate for the customer. This step can be fairly complex as in a perfect scenario, we want to calculate the expected revenue of the customer for each of these time periods, estimate the cost of servicing the customer including product costs and customer service costs. In reality, we almost always need to simplify the calculation into something which is not too complex but still serves the business objective. We will in this section focus on estimating expected revenue, which is the most usual way of estimating customer value. i.e. predict expected revenue.

Expected revenue prediction itself can be fairly complex and consists of estimating (a) revenue from existing products the customer buys/subscribes (b) increase in revenue due to adding more products. Again this approach might not be relevant for all businesses as they might not have additional products, so part (b) might not be relevant; or the customers purchase might be a basket (grocery) and in which case it might make more sense to just predict the dollar volume of the basket over time. We will discuss both part (a) and (b) here.

Estimating Revenue from Existing Products

There are several approaches one can take here. Here is a listing:

- Time series (PROC FORECAST)
- Linear Regression (PROC REG)
- Generalized Linear Model (PROC GENMOD)
- Ordinal (Proportion Odds) Logistic Regression (PROC LOGISTIC)

Time Series is probably the most simplistic way of trying to predict the future values of customers expected revenue. While there are fairly sophisticated time series approaches available is SAS (ex. PROC ARIMA), at the customer level we are looking to at a few hundred thousand to several million individual forecasts. In this situation, it is impossible to generate individual models manually, and automated forecasting techniques need to be used. PROC FORECAST is a very good option for generating these forecasts using these forecasts.

However, as appealing as automated time series forecasts appear to be, from our experience it is not the best way to generate customer level forecasts. For one thing these techniques depend on past patterns and at the individual level these patterns can have a lot of volatility and noise, making the forecasts unreliable. In addition, by depending on past patterns rather than the drivers of the patterns, these techniques will take much longer to adjust to changes in the primary driver, than an approach that directly incorporates them. This is not to say that time series approaches cannot account for these drivers directly, but it is very difficult to implement these approaches at the individual customer level.

Linear Regression (PROC REG) is a good and simple way to estimate future revenues, and allows the modeler to accommodate drivers directly. However, more often than not revenue is not normally distributed but is rather left skewed with often a very long tail. Trying to fit a normal distribution leads to a drop in the accuracy of the prediction.

To correct for this problem, often a generalized linear regression model (PROC GENMOD) is used. PROC GENMOD allows the modeler to fit distributions other than normal (ex. Poisson) and better fit the revenue distributions. However, some of the nice features related to variable selection are not available in PROC GENMOD and so the model development exercise is very cumbersome. A solution often used is
to develop preliminary models using PROC REG to identify the predictors and then develop the final model in PROC GENMOD using the appropriate distribution.

All the approaches discussed earlier, especially the regression techniques share the flaw that they tend to do worse at the tails. This could lead to the model performing poorly for some groups of customers. An approach taken to overcome this is to predict revenue buckets rather than actual revenue directly. This often makes sense, as business decisions are more often than not made on the basis of cutoffs and buckets, and the actual revenue value is less important than the bucket the customer lies in. Therefore, an approach that gets the bucket estimation more correct is often of greater value even if it does worse in terms of overall revenue accuracy.

The revenue bucket prediction approach requires us to create ordinal groups of the revenue. This can be based on the data frequencies, or business requirements or a combination of both. Once the buckets are created, they can be estimated using an ordinal logistic model (proportional odds model) using PROC LOGISTIC. This approach provides a common estimate for all the levels but varies the probability of a customer being in the different buckets by estimating a different intercept for each bucket.

While each of these approaches have their merits and might be appropriate in different scenarios, in our experience the ordinal logistic approach works very well in most situations and provides the business with actionable differences of the expected revenue from the customer.

**Estimating Revenue from New Products**

As we mentioned earlier in many situations new products might not be relevant for the business scenario. In other situations, a conscious decision might be made to not include this step due to the complexity of the process. In still other situations a staggered approach might be adopted where a simpler implementation of CLTV is initially included and then a more complex measure including this step is added later.

In this stage, we are looking to identify the additional products the customer might buy. This can be done using the following types of models

- Binary Logistic to model Purchase/No purchase (PROC LOGISTIC)
- Multinomial Logistic to model selection across various choices (PROC CATMOD, PROC LOGISTIC)

The multinomial model works well if the customer can choose across several options and no purchase can be added as an option.

A lot of cross sell/upsell is often done with incentive campaigns. While it is not directly needed to calculate CLTV, it is often good to know what the true return from the campaign is. To do so we need to calculate true returns from the campaign and account for the fact that some people would have signed up for the product anyway. This can be done using an approach called Net return or Uplift Modeling.

Here are some high level tips on implementing an uplift model in SAS:

- Combine the population which was given the incentive and the population which was held out
- Create a response variable which classifies anyone who signed up as 1 and other as 0
- Next treat the independent variables in the following way:
  - Create an indicator “I” for incentive. Records that got incentive offer have I=1 others 0
  - For every variable considered for modeling (x’s), create interaction variable I*x
- Build a logistic model with the response variable as target, Use x’s, I and I*x’s and then estimate the model.
- Here is how we use the model to get the net response probability (prob response – prob of natural growth) for any customer we are interested in
Score the customer using the logistic equation assuming I=1
Score the customer using the same logistic equation, but assuming I=0
Take the difference of these probabilities to get the true probability of response

CONCLUSION

Estimating the Life Time Value of a customer is very important exercise and helps a business make the right decision in terms of its marketing and customer retention efforts. However, this is a fairly complex process and each step needs to be well thought out for a strategic, analytics and deployment perspective. SAS offers a lot of alternatives for each of the steps and a good approach accounts for the merits of each before deciding on the final process. A good solution is often the one which is not the “coolest” but one which accommodates the business constraints and builds the best solution possible within these boundaries.

REFERENCES

Mehta Krishna et. al. *Implementing Multiple Imputation In An Automatic Variable Selection Scenario*, NESUG 2007
Mehta Krishna et. al. *Selecting the Appropriate Outlier Treatment In Common Industry Applications*, NESUG 2007

ACKNOWLEDGMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

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

Krishna Mehta
Thinkalytics
Email:kmehta@thinkalytics.com
Web: www.thinkalytics.com