Advanced Business Analytics - Power of Predictive Modeling 226161-0131

Adrianna Wołowiec

e-mail: *awolow1@sgh.waw.pl*

Data-mining approach

Probabilistic models for CLV

Earlier we discussed probabilistic models for CLV. Each one began with a set of assumptions that characterize the relationship between the customer and the organization.

For example, customers join and generate some fixed cash flow until they cancel and never return; the chance that a customer cancels is constant over time and customers; the event that a customer is retained in one period is independent of the event in other time periods.

Based on such assumptions, we could derive an expected CLV.

Data-mining approach

Today we explore an alternative data-mining approach that begins with data rather than assumptions about the customer relationship.

The value of a customer in some future period is modeled directly as a function of what is known prior to the future period. We seek a function that, above all else, fits the data well, and the quality of the fit on an independent holdout sample will be the top priority instead of the assumptions and mathematical model characerizing the relationship.

The data-mining approach is also used to predict response to a single contact.

Two classes of approach

Data-mining approach is concerned with two behaviors.

The first is responding to a single contact, as measured by the revenue attributed to the contact. Such models are often called scoring models.

The second set of behaviors is long-term activity, as measured by the total revenue over an extended period of time in the future. We call this long-term revenue (LTR). Multiplying LTR by a margin and deducting relevant costs gives as estimate of a customer's long-term value.

Two above mentioned classes of models can be combined to estimate the value of contacting an individual customer.

Proxy variable

The data-mining approach to modeling begins by identifying a proxy for the behawior that has already been observed. The proxy is denoted by y.

For example, if we want to estimate which customers will respond to a future contact point, a proxy is how customers responded to a similar contact point in the past.

If we want to predict how customers will respond to an e-mail offer that will be sent tomorrow, we can study a test of the e-mail offer that was mailed out last week.

In the case of LTR, if we want to know how much customers will spend next year, a proxy would be how much the customer spent in the previous year.

Modeling

Having identified a proxy behawior (y), we use the information available prior to the proxy period of time ($\boldsymbol{x}$) to predict the proxy behawior with a data-mining model, denoted by function $\mathrm{f}(\boldsymbol{x})$.

The y variable is called a dependent variable by statisticians, and an output variable by data miners. Statisticians call the $\boldsymbol{x}$ variables independent or predictor variables and data miners might call them inputs or feature variables.

Data miners call the task of creating $\boldsymbol{x}$ variables feature extraction, where the variables are extracted from a relational database. The quality of the feature variables is one of the most important determinants of the predictions.

Modeling

We can think of the data being generated as follows:

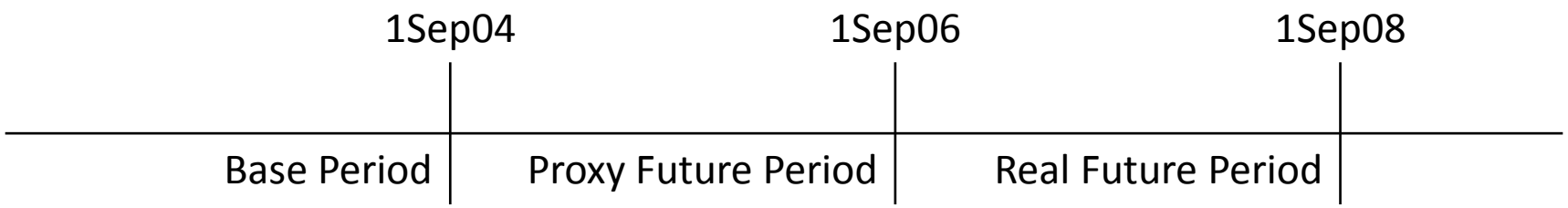
$$y = f(\boldsymbol{x}) + e,$$

where e is an error, usually assumed to be homoscedastic and independent across observations.

We then apply the model to the current information to predict the behavior.

Example 1

The last example in the previous class showed how to use a recency-only migration model to estimate the customer equity of all donors during the two-year period between September 1, 2006 and August 31, 2008. This example shows how to use a data-mining approach for the same problem.



Solution

The first step is to find a proxy for the set of behaviours that we want to model. We use the two-year period from September 1, 2004 and August 31, 2006 as a proxy for the real period. For each customer, we sum all donations during the future period so that we have a measure of long-term (two-year) revenue (the *ltr* variable in the code), which corresponds to the dependent variable y. In the code we use the retain statement to roll up transaction files

Comparison

There has been a long and lively debate about the merits and limitations of the two approaches – probabilistic and data-mining approach. It is important to note some of the key points.

First, extrapolation beyond the end of the future period is especially questionable. Our models will provide an estimate of long-term value over three years, and no more.

Of course, probabilistic models make assumptions that may or may not be true. Many of them are difficult to confirm. When the assumptions made by model are untrue then the conclusions drawn from it are suspect.

Regression models for highly skewed data

Two-step models

Introduction

The distribution of the dependent variable for LTR and scoring models is usually bimodal. One mode consists of zeros, indicating that a customer does not make any purchases in the future period. The right mode is usually approximately lognormal.

Such a dependent variable is problematic for linear regression because the variables predicting whether a customer will respond are different from those predicting how much a customer will spend if the customer responds.

Combining these separate behaviors – response and spend conditional on response – creates modeling difficulties. One approach to modeling such variables is to use a two-step model, which estimates two separate models and combine the predicted values.

Models in two-step approach

1. **Response model**. Estimates the probability that the customer will respond to the contact, that is, $y = 0$ versus $y > 0$. Recency and frequency variables tend to be the most important predictors of future response. Sometime a purchase rate, frequency/time on file, is used in place of frequency, because customers with longer tenure have had more opprotunities to purchase.
2. **Conditional-spend model**. For the responders only, estimate the amount spent or donated. The multiplicative model is often used fot this task. Monetary value, perhaps normalized by frequency to give an average order amount or time on file to give an average amount per period, is usually the best indicator of future spend for those who are active.

Predictions

After both models have been estimated, the predictions from the two models are multiplied to give the final score. Let R be a random variable indicating whether a customer responds with probability $P(R = 1) = \pi$. We ultimately want to know the expected future spend, $E(Y)$, which can be decomposed as follows:

$$E(Y) = E[E(Y|R)] = (1 - \pi)E(Y|R = 0) + \pi E(Y|R = 1) = \pi E(Y|R = 1)\,,$$

because the expected spend of someone who does not respond is 0.

Let π_i be the probability that customer i responds. Estimate the following model with logistic regression:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{a}\,,$$

where $\boldsymbol{a}$ is a vector of slope coefficients.

Predictions

The estimated probability of response is

$$\hat{\pi}_i = \frac{1}{1+e^{(-\boldsymbol{x}_i'\boldsymbol{a})}}$$

Using only observations where $y_i > 0$, estimate a linear regression predicting $\log(y_i + 1)$ from the predictor variables producing slope vector $\boldsymbol{b}$ and mean squared error S_e^2. Final scores for a future customer are nonlinear functions of the covariates $\boldsymbol{x_0}$ given by:

$$\hat{y}_0 = \frac{e^{(\boldsymbol{x_0'b}+\frac{S_e^2}{2})}-1}{1+e^{(-\boldsymbol{x_0'a})}}$$

The exponent in the numerator converts the log-amount used as the dependent variable in the regression inot raw amounts.

Example 2

Continue the previous example and estimate a two-step data-mining model with RFM as predictors.

Evaluating data-mining models

Introduction

Suppose we have an estimate $\hat{f}$ of function f. We need to evaluate the quality of the estimated model.

Here we will cover two issues:

- devising metrics to evaluate the model
- estimating the metrics

Two families of metrics are commonly used in practice. One masures the typical size of the errors e and the other estimates revenues realized by using $\hat{f}$ to select a certain numer of people to receive a contact point (for example, gains tables).

Both families of metrics should be estimated on a test sample of data rather than the training sample used to estimate $\hat{f}$.

Residual-based metrics

The quality of the fit is evaluated through the residuals $\hat{e}_i = y_i - \hat{f}(\boldsymbol{x_i})$. There are many metrics that are some function of the $\hat{e}_i$ values.

The sum of squared errors is:

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} [y_i \ - \hat{f}(\boldsymbol{x_i})]^2$$

Often we rescale SSE so that it measures the average squared residua or mean squared error ($MSE = SSE/n$). MSE is also called the averagesquared error (ASE).

The root mean squared error, $RMSE = \sqrt{MSE}$, measures the typical size of a (not squared) residua.

Residual-based metrics

Another rescaled version of SSE is the coefficient of determination, $R^2 = 1 - \frac{SSE}{SST}$, which is the fraction of variation in y explained by the model.

The constant $SST = \sum_i (y_i - \bar{y})^2$ is called the total sum of squares, giving the sum of squared residuals from the baseline intercet model $\hat{f}_0(\boldsymbol{x_i}) = \bar{y}$ (for all i).

We can think of R^2 as making a comparison between our model $\hat{f}$ and the simplest possible model consisting only of an intercept, whic his constant over all values of $\boldsymbol{x}$ and does not model any dependence of y on $\boldsymbol{x}$.

Gain tables

For many applications, having small residuals is not the primary objective. For example, a model that predicts the response to some marketing contact point will probably be used to select which customers will receive the contact. The ultimate objective in this situation is to maximize revenues or profit, and it would be desirable to have a metric that directly matches this objective. Gains tables give the answer.

Gains tables

In practice, the predicted values from the regression function $\hat{f}(\boldsymbol{x_i})$ will be sorter. The ordering will determine which customers receive the contact point, with customers having larger predicted values receiving the contact before those with smaller values.

Gains tables are created as follows:

1. Find the deciles of the predicted values. Quantiles other than deciles could also be used, but deciles are probably the most commonly used quantile in practice.
2. Compute actual reveue by decile.
3. Compute cumulative columns.

Example 3

Create gains tables for the earlier estimated regression model.