

MEASURING REAL-TIME PREDICTIVE MODELS

SAMUEL STEINGOLD, RICHARD WHERRY, AND GREGORY
PIATETSKY-SHAPIO

ABSTRACT. In this paper we examine the problem of comparing real-time predictive models and propose a number of measures for selecting the best model, based on a combination of accuracy, timeliness, and cost. We apply the measure to the real-time attrition problem.

1. INTRODUCTION

Comparing quality of different models is an important practical task. When the model deals with static data, this is a well-established area with many known model measures, like Lift, Response Ratio, L-quality etc (see [1], [2] and [3]).

Recently data miners have begun to deal with dynamic and real-time data. We can use the latest transactions to predict how likely the customer is to buy this product, click on that link, or disconnect the service. When using real-time data, in addition to traditional dimensions of accuracy and cost, we also have to compare models on the time dimension. If model A makes the same correct prediction as model B , but earlier, then model A should be preferred.

We have recently conducted extensive experiments with building real-time attrition models and in this paper we summarize our experiences.

Suppose a company has a set C of customers and a continuous transaction stream X which, possibly together with historical data, is used to predict attrition of the customers. The model should be trained using the data up to the 1st day of the test period (usually a month), and then run daily on the set of transactions up to that day (and, possibly, on all the other historical data accumulated so far). The model should return the probability that the customer will attrite before the end of the period.

Date: June 15, 2001.

Key words and phrases. Real-time, Database marketing, measurement, predictive models.

The difference from the usual setup is that it is important to identify the potential attritors as early as possible, so that some actions can be taken to retain the customer.

2. DESIRABLE PROPERTIES OF THE MEASURE

A model $M(c, t)$ is a function of two arguments: the customer $c \in C$ and time $t \in [0; T]$, which means that the actual model scoring algorithm gets all the historical and demographical data for c as well as all transactions $x \in X$ up to time t . M returns (an estimate of) the probability $p \in [0; 1]$ of the predicted event (“attrition”) by the end of the time period (say, a month, in which case $T = 30$). $M(t)$ is the model at time t . The boolean value $A(c)$ says whether the customer c will actually attrite by the end of the month.

As specified in the Introduction, the model which identifies the potential attritors early is better, thus, the model quality $Q(M)$ should satisfy the following property: if M_1 is “better” than M_2 in the following sense: for all t

$$(1) \quad \begin{array}{l} \text{if } A(c) = 1 \text{ then } M_1(c, t) \geq M_2(c, t) \\ \text{if } A(c) = 0 \text{ then } M_1(c, t) \leq M_2(c, t) \end{array}$$

then $Q(M_1) \geq Q(M_2)$. In other words, if the model M_1 predicts *true* attritors *earlier* and *non-attritors later* than model M_2 , then it is “better”.

One of the simplest measures would be the L-quality, as defined in [3]. If we define $L(M, t) := \text{L-quality}(M(t))$ to be the L-quality of the model M as run at time t , one can easily see that if the model M_1 is better than the model M_2 in the sense of 1, then for all t we have $L(M_1, t) \geq L(M_2, t)$, so a simple measure like $Q(M) := \sum_t L(M, t)$ would satisfy the requirement. Unfortunately, this measure does not capture the notion of *how much earlier* one model identifies the attritors than the other.

Thus, we need to generalize the current model quality measure to take the time aspect into account. While ultimately the cost/benefit matrix of the (in) correct prediction should be the deciding factor, in many cases it is not known, or subject to change. Thus we start with considering the value-free measures, based on time and accuracy, and will add the value consideration later.

3. VALUE-FREE MEASURES

3.1. Derivation. For a customer $c \in C$ who is an attritor ($A(c) = 1$), consider the following measure:

$$(2) \quad Q_0(c) = \frac{2}{T^2} \int_0^T M(c, t)(T - t)dt$$

When the model stably predicted that this customer would attrite (i.e., $M(c, t) = 1$ for all t), we have $Q_0(c) = 1$, and if the model never predicted attrition ($M(c, t) = 0$ for all t), then $Q_0(c) = 0$. If the model started to predict attrition at time $0 < \tau < T$, i.e.,

$$(3) \quad M_\tau^1(c, t) = \begin{cases} 0 & \text{when } t \leq \tau \\ 1 & \text{when } t > \tau \end{cases}$$

then

$$(4) \quad Q_0(c) = 1 - \frac{\tau}{T}$$

Thus Q_0 measures how much time the company would have to contact the potential attritor. Note that, since $T - t \geq 0$ when $t \in [0; T]$, this measure satisfies the requirement 1.

For a customer $c \in C$ who is not an attritor ($A(c) = 0$), we can write

$$(5) \quad Q_0(c) = -\frac{1}{T} \int_0^T M(c, t)dt$$

to charge a penalty for falsely predicting attrition. If the model stopped predicting attrition at time $0 < \tau < T$, i.e.,

$$(6) \quad M_\tau^0(c, t) = \begin{cases} 1 & \text{when } t \leq \tau \\ 0 & \text{when } t > \tau \end{cases}$$

then

$$(7) \quad Q_0(c) = -\frac{\tau}{T}$$

Thus, for an “always positive” model $Q_0(c) = -1$, while for an “always negative” one $Q_0(c) = 0$. Therefore Q_0 measures how long the model gave a false prediction, i.e., the chances the company had to contact the non-attritor. Note that this measure satisfies the requirement 1.

Combining formulas 2 and 5 using A and $1 - A$ as a partition of unity, we arrive at

$$\begin{aligned} Q_0(c) &= \frac{2}{T^2} \int_0^T M(c, t) \left(A(c)(T - t) - (1 - A(c))\frac{T}{2} \right) dt \\ &= \frac{1}{T} \int_0^T M(c, t) \left(2A(c)\left(1 - \frac{t}{T}\right) - (1 - A(c)) \right) dt \end{aligned}$$

or

$$(8) \quad Q_0(c) := \frac{1}{T} \int_0^T M(c, t) \left(3A(c) - 1 - A(c) \frac{2t}{T} \right) dt$$

It only remains to average Q_0 over all customers to arrive at a real-time model quality measure:

$$(9) \quad Q_0(M) := \frac{1}{NT} \sum_{c \in C} \int_0^T M(c, t) \left(3A(c) - 1 - A(c) \frac{2t}{T} \right) dt$$

where $N = \#C$ is the total number of customers.

Since both 2 and 5 satisfy the requirement 1, this combined measure satisfies it too.

3.2. Rationale. We want the quality of the similar models 3 and 6 to look similar, namely like 4 and 7 respectively, and formulas 2 and 5 are the only solutions.

Formulas 4 and 7 let us interpret Q_0 as the mean “good time” (the advance warning the model gives us for the true attritors) minus mean “bad time” (the advance warning the model gives us for the non-attritors).

Note that we did not consider false positives (M_τ^1 when $A(c) = 0$) and false negatives (M_τ^0 when $A(c) = 1$). The reason is that both of these situations are highly unlikely in reality: usually the quality of prediction improves with time, i.e., we expect that $|M - A|$ decreases with time. A more realistic examples of false positives are M_τ^1 when $A(c) = 1$, and false negatives are M_τ^0 when $A(c) = 0$, but τ is close to T , i.e., the correct prediction comes too late.

3.3. Examples. Let us compute this measure for some specific models.

3.3.1. “Always Negative” Model. If $M(c, t) = 0$ for all $c \in C$ and $t \in [0; T]$, then obviously $Q_0(M) = 0$.

3.3.2. “Always Positive” Model. If $M(c, t) = 1$ for all $c \in C$ and $t \in [0; T]$, then let $N_a = \#\{c : A(c) = 1\}$ be the number of attritors and $b = \frac{N_a}{N}$ be the base rate, and compute

$$\begin{aligned} Q_0(M) &= \frac{1}{NT} \left(\sum_{A(c)=1} \int_0^T (2 - \frac{2t}{T}) dt + \sum_{A(c)=0} \int_0^T (-1) dt \right) \\ &= b - (1 - b) \\ &= 2b - 1 \end{aligned}$$

3.3.3. *“Perfect” Model.* If $M(c, t) = A(c)$ for all $c \in C$ and $t \in [0; T]$, then

$$\begin{aligned} Q_0(M) &= \frac{1}{NT} \sum_{A(c)=1} \int_0^T \left(2 - \frac{2t}{T}\right) dt \\ &= b \end{aligned}$$

3.3.4. *“Random” Model.* If $M(c, t) = b$, where b is the base rate, as above, for all $c \in C$ and $t \in [0; T]$, then this is just a multiple of the “always positive” model and $Q_0(M) = b(2b - 1)$.

3.3.5. *Summary.* Here is the summary for the four models:

Always Positive	$2b - 1$
Always Negative	0
Perfect	b
Random	$b(2b - 1)$

3.4. **Discussion.** Figure 1 shows that the “Perfect” model is the best, and its advantage is the highest when the base rate is $\frac{1}{2}$. When the base rate is 1, “Perfect”, “Random” and “Always Positive” models are identical, and, indeed, their Q_0 is the same – one. When the base rate is 0, “Perfect”, “Random” and “Always Negative” models are identical, and, indeed, their Q_0 is the same – zero.

As expected, the “Random” model’s quality is between that of “Always Positive” and “Always Negative” ones, and “Always Positive” wins when the base rate $b > \frac{1}{2}$ and loses when $b < \frac{1}{2}$.

It is better to normalize Q_0 so that the “Perfect” model always has $Q = 1$, while the “Random” model has $Q = 0$, i.e.,

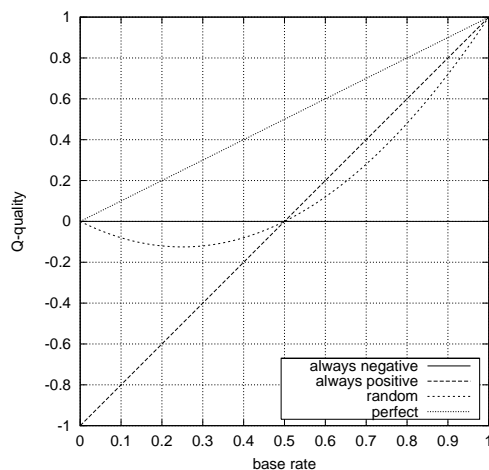
$$Q(M) = Q_n(M) = \frac{Q(M) - b(2b - 1)}{b - b(2b - 1)}$$

This means

$$(10) \quad Q(M) := \frac{1}{2NTb(1 - b)} \sum_{c \in C} \int_0^T (M(c, t) - b) \left(3A(c) - 1 - A(c) \frac{2t}{T}\right) dt$$

Since the original measure 9 satisfies the requirement 1, this measure satisfies it too.

Then we would have

FIGURE 1. Comparison of Q_0 for various models

Always Positive	$(2b - 1)/2b$
Always Negative	$(1 - 2b)/2(1 - b)$
Perfect	1
Random	0

Note that the graphs on figure 2 are symmetric.

4. VALUE-BASED MEASURES

Business practice has shown that any data-mining effort should take the customer value into account, otherwise valuable resources would be wasted on retaining an unprofitable customer whose probability of attrition is high, while no effort would be made to retain a profitable customer with low attrition probability.

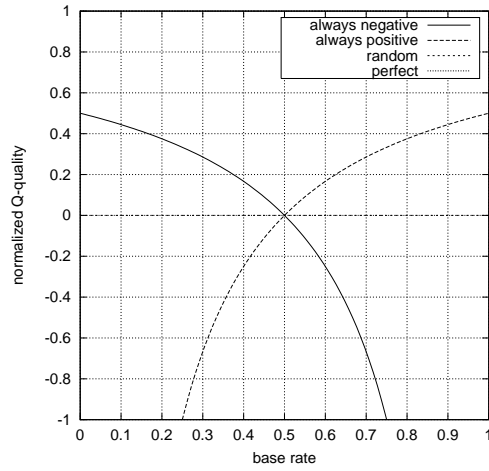


FIGURE 2. Comparison of $Q = Q_n$ for various models, normalized

Thus we must ascribe a customer a value $V(c)$, take it into account during model training and evaluate the model using

$$(11) \quad Q(M) := \frac{1}{2NTb(1-b)} \sum_{c \in C} \int_0^T (M(c, t) - b) V(c) \left(3A(c) - 1 - A(c) \frac{2t}{T} \right) dt$$

Actually $V(c)$ should be not the past value of the customer, but his projected total future value.

Again, this measure satisfies the requirement 1 when the value V is non-negative, since

$$\begin{aligned} V(c) \left(3A(c) - 1 - A(c) \frac{2t}{T} \right) &\geq 0 \quad \text{when } A(c) = 1 \\ V(c) \left(3A(c) - 1 - A(c) \frac{2t}{T} \right) &\leq 0 \quad \text{when } A(c) = 0 \end{aligned}$$

5. REAL WORLD EXAMPLES OF REAL-TIME MODELS

5.1. Problem Description. In this section, we will evaluate the performance of two models. The goal of each is to identify those customers that exhibit behavior that leads to attrition. Once these customers are identified, action should be taken on behalf of the institution to prevent attrition. This usually takes the form of a marketing offer.

It is assumed the attrition models will be part of a data mining system that operates on transactions in real-time (e.g., customers are scored at each transaction). It is desirable to make such a system adaptive to transactions as they happen in real-time. Such a system should have a method of measuring the effectiveness of a deployed model. One way to do this is using the usual lift, accuracy, or l-quality. However, these measures are static and do not capture the dynamic nature of the production environment. Model quality, developed in the previous sections, provides an excellent measure for such a system.

5.2. Data. The data that we use is a combination of historical and transaction banking data. We have two months of historical data (January and February) and one month (March) of transaction data. We are trying to predict attrition in the months of March and April. We will compare two models and judge their performance against each other and to that of a random model using the model quality measure developed in the section 3.

The data is split at the customer level into a training and held-out datasets so that *all* transactions belonging to a certain customer are in either the training dataset or *all* transactions for that customer are in the held-out dataset.

The training takes place over the entire period of the 31 days of March. The held-out data is divided into 5-day increments (March 01-05, March 06-10, March 11-15, March 16-20, March 21-25, and March 26-31). This will allow us to estimate $Q(M)$.

5.3. Estimating Model Quality. To estimate $Q(M)$, we use the held-out set C_h of customers, which was not used in training the model, and change the order of summation and integration in the equation 10. We also replace the integral with a sum, computed for t_1 being March 5, t_2 – March 10, t_3 – March 15, t_4 – March 20, t_5 – March 25, and t_6 – March 31. In addition, we split the sum over customers in 10 into a sum for attritors and a sum for non-attritors. This will lead to an efficient computation of $Q(M)$ with the assumption that the number

Period	$Q_0(M)$	$Q_n(M)$ (Normalized)
Model 1	-0.0555	-0.8712
Model 2	-0.0013	0.0526
Random	-0.0282	0.0

TABLE 1. The quality measures for the experimental models

of non-atritors is much greater than the number of attritors. We get the following expression:

$$(12) \quad Q(M) := \frac{1}{2N_h T b(1-b)} \sum_{t=1}^{t=6} \left[\sum_{c \in C_h, A(c)=0} (b - M(c, t)) + \sum_{c \in C_h, A(c)=1} 2(M(c, t) - b) \left(1 - \frac{t}{T}\right) \right]$$

Here $N_h = \#C_h$ is the number of customers in the held-out set, and $M(c, t)$ is the model score computed for the transactions of the customer c up to the time t plus his historical and demographical data (if the model uses it). If the model uses only the transaction data and the customer had none in $[0; t]$, then the base rate $M = b$ is used (which is obviously equivalent to ignoring such customers). Note that b is the global base rate across the whole customer set C , which should be the same as the base rate in the held-out set C_h .

Unfortunately, the inner sums in equation 12, do not make much sense in themselves, in particular, they do *not* estimate $Q(M)$ restricted to $[0; t]$ (instead of $[0; T]$).

5.4. Results. Both models were trained using the same dataset and are evaluated on the same held-out dataset. The training dataset consists of transactions for 6,814 customers. The held-out dataset consists of transactions for 3,440 customers. The average number of transactions per customer is approximately 18. The held-out data transactions were divided into each of the six periods ($t = 5, 10, 15, 20, 15, 31$). For the first period, March 1-5, there were 1,865 customer that performed transactions. By the end of March 3,440 customers had performed transactions.

Table 1 summarizes the results.

With a base-rate of 0.03, both models perform poorly. Notice that for both Model 1 and Model 2, $Q_0(M) < 0$. This tells us that for both

models, the mean “bad time” is greater than the mean “good time”. Accordingly, both models are giving more weighted advanced warnings for non-attritors than for attritors. The fact that $Q_0(M) < 0$ is not necessarily bad. In fact, $Q_n(M_1) > 0$ indicates that Model 1 performs better than the random model.

5.5. Real-Time Environment. We have evaluated the models on the held-out dataset. Model 1 performs better than Model 2, so we choose to deploy Model 1 to a real-time environment. Once the model is deployed, it will be monitored. The model quality, $Q(M)$ developed in the previous sections provide an excellent way to measure the performance of the model.

However, in order to use $Q(M)$ to measure the performance, we need to choose T , the period over which the model will be evaluated, and select the held-out subset of customers whose transactions will be used to estimate $Q(M)$.

If the deployed model begins to perform below certain level, we can re-build and re-deploy the model. Alternatively, if the learning algorithm supports it, we can adapt the model to the new information.

6. DISCUSSION

6.1. Cross-Sell Models. The models developed in section 5 were attrition models where the goal was to identify attritors as early as possible. In a cross-sell model, the goal is to identify which product, if any, is the best sell for the current customer, based on their behavior and likeness with customers that currently own the product. For example, a bank may wish to sell certificates of deposits (CD) to its current base of customers. It may be desirable to identify these potential customers as early as possible. Continuing with the CD example, a customer may be gradually building up a balance in their savings or checking, along with other indicative behavior. The bank may want to intervene as early as identifiably possible in this process and make an offer of a CD to the customer. As in the case of attrition, a static measure will measure which model is better at time t , but not provide you with enough information as to which model made the correct prediction earliest. Model quality, $Q(M)$, can be used in this situation.

6.2. Customer Value. It bears repeating the dictum of section 4 that the reason for data mining in the business setting is to increase the value of the customer to the company and vice versa. Therefore it is

crucial to use the (reasonable estimate of the) customer value in the formula 11 if at all possible.

REFERENCES

- [1] Foster Provost, Tom Fawcett “Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions”, *Proceedings of KDD-97*, pp. 43-48, AAAI Press, 1997.
- [2] Gregory Piatetsky-Shapiro, Brij Masand “Estimating Campaign Benefits and Modeling Lift”, *Proceedings of KDD-99*, pp. 185-193, ACM Press, 1999.
- [3] Gregory Piatetsky-Shapiro, Sam Steingold, “Measuring Lift Quality in Database Marketing”, *SIGKDD Explorations*, Vol. 2:2, (2000), 81-86

SAMUEL STEINGOLD, XCHANGE INC, (617) 790-2522

E-mail address: `sds@exchange.com`

RICHARD WHERRY, XCHANGE INC, (617)-790-2542

E-mail address: `rwherry@exchange.com`

GREGORY PIATETSKY-SHAPIRO, XCHANGE INC & KDNUGGETS, (617) 232-7512

E-mail address: `gps@kdnuggets.com`