Distance to Event vs. Propensity of Event
A Survival Analysis vs. Logistic Regression Approach

Abhijit Kanjilal
Fractal Analytics Ltd.

Abstract:
In the analytics industry today, logistic regression is a very robust and proven technique to predict the propensity of certain events. As a result, it has been very popular in marketing analytics. Given the same marketing problems, a survival analysis tries to answer them slightly differently – it predicts a (time) distance to an event, instead of the propensity of event. With an objective to access the predictability, robustness, and strength of survival analysis approach as compared to logistic regression, we carried out an analysis in the context of an Activation Management Program. The results clearly show that one survival analysis model is as good as multiple logistic regressions for different prediction window, in terms of predictability and strength.

When we talk about marketing activity in a customer’s lifetime, it probably starts with activation and ends with attrition. These events basically constitute the (obviously closed) boundaries of the premise of marketing. Hence effective marketing around these events is extremely crucial for organizations to drive a significant profitability. This demands the prediction of such events and definition of strategies based on a customer’s inclination towards these.

The entire challenge becomes measuring a customer’s inclination towards marketing events. This can be done in two ways, a) through a Propensity of Event approach or b) through a Distance to Event approach.

When we are talking about the propensity of the event approach, it basically scores a probability of the event for each customer. This approach inherently assumes a prediction window in which customers will be assigned a probability of the event; for example, the activation probability of newly acquired customers in the next 12 months. Given this example, we generally would love to build a logistic model. Now, think about the customers who are likely to activate within first six months of on book and customers who are likely to activate within 1-12th month on book. The profiles of these two sets may differ, hence prediction from P6 (activation probability in 6 months) and P12 (activation probability within 12 months) may not be comparable. To understand the profiles of these two sets we basically need to build two models, one with prediction window of six months, and another of 12 months. If we divide the window further, we end up building more models.

Do we have any technique so that we get the solution at a go? The answer is survival analysis. Medical science has been using this technique extensively. We should keep in mind that medical science is very sensitive towards errors in prediction. If marketing is also very sensitive towards errors in predicting “right customers at right time with right offers,” then it makes sense to explore the technique. Does it add any value to the process? This gives us the motivation to access survival analysis as compared to logistic regression in the context of Activation Management Program.

Theoretical Model:
Logistic regression is a model used for prediction of the probability of occurrence of an event. In this case our target variable is of binary type – e.g. attrited or not, or card activated or not. It makes use of some predictive variables, either numerical or categorical, that contain all available information about the objects.
logit($p_i$) = ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1,i + ... + \beta_k x_k,i

where $p_i$ is the probability of the event for the $i^{th}$ object and x’s are the risk factors affecting the event. $\beta$’s are the parameters of the model. Hence the probability boils down to

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1,i + ... + \beta_k x_k,i)}}$$

In logistic regression, we are interested in studying how risk factors are associated with the target event. Sometimes, though, we are interested in how a risk factor affects time to event. Survival analysis is used to analyze data in which the time until the event is of interest. The response is often referred to as a failure time, survival time, or event time.

The beauty of survival analysis is that it can tackle the problem of censored data. Censoring is present when we have some information about a subject’s event time, but we don’t know the exact event time.

There are several survival analysis techniques, Cox Proportional Hazard model is the most popular one. In this research we only focused on this technique.

Let $T_j$ denote the time of the event. Our data, based on a sample size n, consists of a triple

$$(T_j, \delta_j, X_j), j = 1,...,n$$

Where $T_j$ is the time on study for the $j^{th}$ customer, $\delta_j$ is the event indicator, and $X_j = (X_{0j}, X_1, X_2,....X_p)$ is the vector of risk factor affecting the distance to event.

Now, $h(t | X) = h_0(t) e^{(X' \beta)}$

where $h_0(t)$ is an arbitrary baseline hazard rate, $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the parameter vector and $e^{(X' \beta)}$ is a known function. Because $h(t | X)$ must be positive, a common model for $e^{(X' \beta)}$ is

$$e^{(X' \beta)} = \exp(\sum_{k=1}^{p} \beta_k X_k)$$

The model estimates will be arrived through the following partial likelihood function maximization

$$L(\beta) = \prod_{j=1}^{n} \frac{\exp(\sum_{k=1}^{p} \beta_k X_k)}{\sum_{j \in R(t_j)} \exp(\sum_{k=1}^{p} \beta_k X_k)}$$

Empirical Model:

Having these two techniques in place, we tried to address Activation Management Program for a bank credit card. Our objective was to access customers’ inclination towards activation within six months after acquisition and based on the results to prioritize promotional strategy and speed up the activation process of customers from their natural pace. Customers’ application information was used to build models.

We divided the six months prediction window in three parts: 0-2 months, 0-4 months, and 0-6 months, and built three logistics regression models to predict two-, four-, and six-month activation probability.

On the other hand we built one survival analysis and computed two-, four-, and six-month activation probability.

To construct this solution, we first fit a proportional hazard model to the data and obtain the partial likelihood estimators.

Let $t_1 < t_2 <.....< t_D$ be the distinct distance to activation and $d_i$ be the no of active customers at $t_i$.

Let $W(t_i; \beta) = \sum_{j \in R(t_i)} \exp(\sum_{k=1}^{p} \beta_k X_{0j})$

The estimator of the cumulative baseline activation rate

$$\hat{H}_D(t) = \int_{0}^{t} h_0(u) du$$

is given by

$$\hat{H}_D(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; \beta)}$$

The estimator of the baseline dormancy function,

$$\hat{S}_D(t) = \exp[-\hat{H}_D(t)]$$

The dormancy probability of a new customer with a given set of Risk Factors $X_0$ will be

$$S_0(t | X_0) = S_0(t) \exp(\sum_{k=1}^{p} \beta_k X_k)$$

Activation Probability = 1 – Dormancy Probability

![Graph showing activation probability over time](image-url)
Activation probability will be a monotonically increasing function in time. Here is the beauty of survival model, to get some kind of time function in logistics regression set up, we would need to build several models.

The results of the analysis have been shown below – one survival model vs. multiple logistic models.

(Results are based on...)

**Proportional Hazard Model**

![Graph showing Activation Probability Across Quartiles](image)

**Proportional Hazard model is clearly able to distinguish the 4 quartiles in terms of activation probability**

**Logistic Regression Model**

![Graph showing Activation Probability Across Quartiles](image)

**Logistic Regression model is clearly able to distinguish the 4 quartiles in terms of activation probability**

*Quartiles are defined based on 6 months activation probability*
In terms of capturing the activated and non-activated customers, the proportional hazard model has performed slightly better than the logistic model.

In terms of capturing the activated and non-activated customers, the proportional hazard model has performed slightly better than the logistic model.

In terms of capturing the activated and non-activated customers, the proportional hazard model has performed slightly better than the logistic model.
Proportional Hazard Model (actual & expected activation rate)

Logistic Regression Model* (actual & expected activation rate)

*Quartiles are defined based on 2 months activation probability
Logistic Regression Model* (actual & expected activation rate)

*Quartiles are defined based on 4 months activation probability

Logistic Regression Model* (actual & expected activation rate)

*Quartiles are defined based on 6 months activation probability
Conclusion:
This case study shows that survival analysis model does as good a job as multiple logistic regression models for a different window in terms of strength and predictability. Survival Analysis differentiates individuals in terms of distance to event, and marketing strategies can be prioritized for optimization to drive higher profitability. The logistic regression model is unable to address the same business problem in similar manner.

It should be noted that the results are based on a limited range of geography and product data. This case-study does not include time-dependent risk factors into consideration. It would be useful to further generalize them in future studies.

Reference:
1. Survival Analysis by John P. Klein, Melvin L. Moeschberger, Springer, 2005
3. Modeling Survival Data: Extending the Cox Model (Statistics for Biology and Health) by Terry M. Therneau and Patricia M. Grambsch, Springer, 2005