MWSUG 2016 - Paper AA21

Use Multi-Stage Model to Target the Most Valuable Customers

Chao Xu, Alliance Data Systems, Columbus, OH
Jing Ren, Alliance Data Systems, Columbus, OH
Hongying Yang, Alliance Data Systems, Columbus, OH

ABSTRACT

To predict the likelihood of customers to be engaged with the business, logistic model is usually employed in the current marketing industry. For example, marketers utilize logistic regression model to predict the probability of customer's response to a marketing offer. As such, only those customers with higher propensity to respond will be selected for targeting to minimize the cost and maximize the return on investment (ROI). However, logistic model does not consider customer's value after response, which should be also important to the business. In this paper, we define the most valuable customers as the ones who have high propensity to respond and to be engaged longer. We combine logistic regression model with survival model to target these most valuable customers. PROC LOGISTIC, PROC LIFETEST and PROC PHREG are explored and utilized in the two-stage model. We compare our predicted results with the results which are generated from traditional logistic regression model. We found that by applying the two-stage model, most valuable customers can be selected and ROI of campaign will also be improved. All programming is executed in the environment of SAS Enterprise Guide® Version 7.

INTRODUCTION

In the marketing industry, gaining new customers, retaining existing customers, and reactivating inactive customers are three major ways to keep and increase active customer base. Generally, "New" customers are defined as those who first get engaged with a business; "Active" customers are those who maintain their engagement; "Inactive" customers are those who have been away from the business for a relatively longer time.

It is common for a business to have a large population of inactive customers. Reengagement campaign, also known as reactivation campaign, is used by many marketers to reach out to those inactive customers who have previously purchased from their business but have become disengaged. Typically, the cost is high if reactivation offers are sent to the entire inactive customer base. Instead, it is better to select a group of customers with high propensity to come back. Logistic regression model is a good way to predict the probability of reengagement.

While the reengagement rate is important, the longevity after customer's reactivation is also very valuable to a business. Thus, increasing reengagement rate and elongating the lifetime of those customers will maximize the overall return on marketing investment. For example, reactivated customers may only have one or two purchase activities after coming back, and then become inactive again. In this case, the survival time is very short compared to the ones who are active for a longer time. The longer they stay, the more revenue they could possibly generate. Therefore, it is best to select the customers with high likelihood to reengage and have long survival time after reengagement.

This paper will focus on the reengagement and longevity of these reengaged customers. We combine logistic regression model and survival model to help selecting the customers who do not only have high propensity to reactive, but are also likely to stay with a business for a longer time. By applying this two-stage model, a business is able to optimize its ROI and increase customer loyalty after reengagement.

METHODOLOGY

LOGISTIC REGRESSION AND ITS LIMITATION

Logistic regression is a statistical model within which the dependent variable is binary. The formula is stated as below:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

Where p is the probability of particular outcomes.

Logistic regression model has been widely used for reactivation campaign. Instead of sending promotions to all customers, in most cases, marketing modelers use logistic regression model to rank customers and select customers who have high response probability. Therefore, those selected top ranked customers have high propensity to respond to reactivation campaign compared with random selection. By increasing reengagement rate, logistic regression model helps a business to invest towards right customers and drive incremental sales.

In this paper, we use logistic regression model as the first stage to predict the likelihood of reactivation. Logistic regression model is commonly used to explain the relationship between one dependent binary variable and several independent variables. For reactivation campaign, the dependent binary variable is whether an inactive customer has repurchased or not after receiving the reactivation offer. We code the dependent variable as 1 – the inactive customer reactivates, 0 – the inactive customer does not reactivate.

We use PROC LOGISTIC to fit the logistic regression model. After this, we have the probability of reactivation of each inactive customer. The higher probability the customer has, the higher likelihood to reactivate. Thus, logistic regression model would be very helpful to select inactive customers with high likelihood to reengage.

When we use logistic regression model to select customers for reactivation campaign, we only consider the likelihood of customer's reactivation. Meanwhile, our goal is not only to have the inactive customers come back, but also to keep these customers active/engaged as long as possible. The longer these reactivated customers stay, the higher revenue a business will have for the win-back campaign. From this perspective, we need to select customers with higher reactivation likelihood and greater longevity as well.

SURVIVAL MODEL

Whether or not reengaged customers generate high ROI also depends on if they stay, purchase, and are loyal to the business after reactivation. It is uncertain that for those reactivated customers, how long they will stay with and be loyal to the business after reactivation. As a result, customer's long-time value should also be considered in addition to the propensity to reengage. Thus, at the time of reactivation campaign, it is more valuable to put investment on customers who have high propensity to respond, have high long-time value, and stay with the business longer.

Survival analysis is known as a powerful tool for modeling factors that influence time of an event. It has also been widely used in clinical field to estimate patient survival rate over time. In recent years, survival model has also been adopted in retail marketing in the prediction of customer's churn, attrition, and tenure. In this paper, we will use survival analysis to assess the number of inactive customers after reactivation per population at risk of attrition per unit time.

Kaplan-Meier Estimator

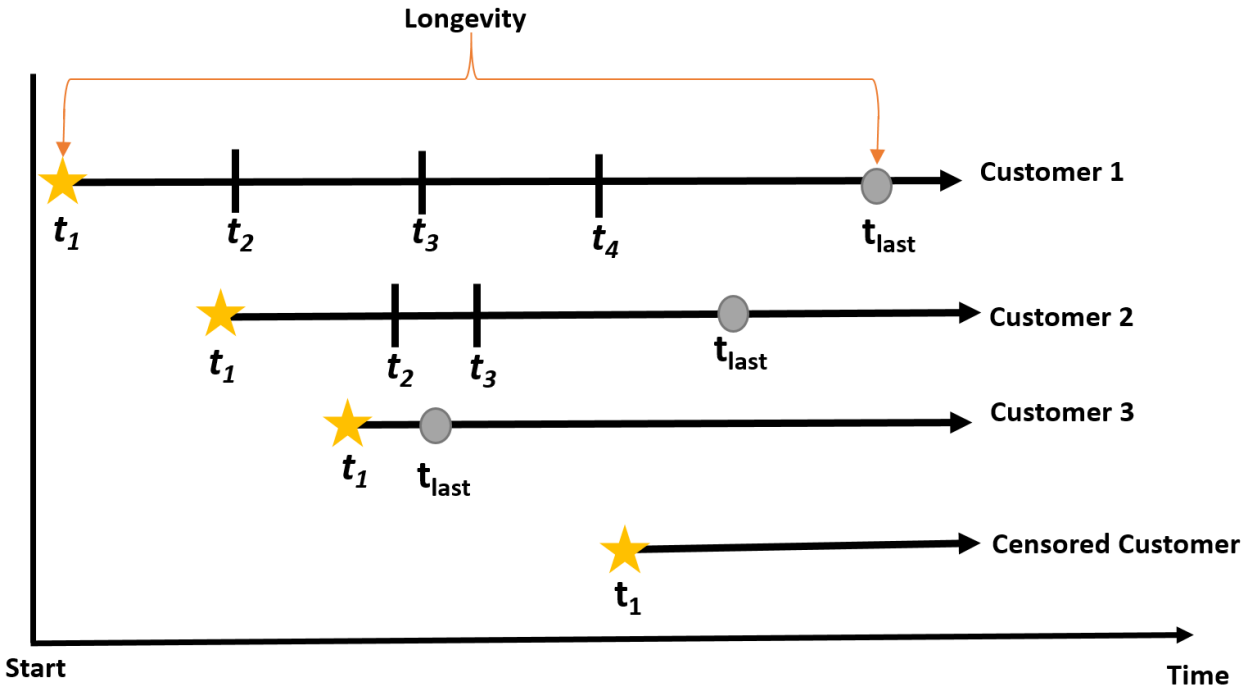

Figure 1. Illustration of Survival Model Setting.

As shown in Figure 1, in our study, customers enter after they make their first purchase (t_1). Event happens when they make their last purchase (t_{last}) before observation end date. Longevity is defined as the time difference ($t_{last} - t_1$) between first purchase date and last purchase date. We define the customers who never purchase or make only one purchase during observation window as censored customers.

We use PROC LIFETEST to examine the pattern of longevity by creating the Kaplan-Meier estimate of the survival function.

The Kaplan-Meier estimator for survival function is calculated as:

$$\hat{S}(T) = \prod_{T_i \le T} \frac{N_i - D_i}{N_i}$$

Here, T_i denotes the longevity of reactivated customers, D_i denotes the number of customers with a longevity of T_i, and N_i denotes the remaining active customers.

Cox Proportional Hazards Model

We use PROC PHREG to perform survival regression modeling. Cox proportional hazard model has been widely used for survival analysis in investigating time-to-event data. In SAS, PROC PHREG has been a powerful tool used for survival analysis, which is based on the Cox proportional hazards model. The longevity of each customer is assumed to follow its own hazard function $\lambda_i(t)$, which is expressed as:

$$\lambda_i(t) = \lambda(t; Z_i) = \lambda_0(t)\exp(Z'_i\beta)$$

Where:

- $\lambda(t;Z_i)$ is the hazard function for the i*th* customer at time t;
- $Z_i = (X_{1i}, X_{2i}, \ldots, X_{ki})$ is the vector of explanatory variables for the i*th* customer;
- $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, i.e., when $X_{1i}=0$, $X_{2i}=0, \ldots, X_{ki}=0$;

- β is the vector of unknown regression parameters that is associated with the explanatory variables. The vector β is assumed to be the same for all individuals.

The hazard ratio $\lambda_i(t)/\lambda_0(t)$ thus can be interpreted as the relative risk of a customer to be inactive again after reactivation at time t.

The log of the hazard ratio can be further expressed as a linear combination of parameters:

$$\log\left(\frac{\lambda(t;Z_i)}{\lambda_0(t)}\right) = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots . + \beta_k X_{ki}$$

The hazard ratio can be expressed as:

$$\lambda(t;Z_i) = \lambda_0(t)\exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots . + \beta_k X_{ki})$$

MODELING PROCESS

In this paper, at the first stage we use logistic regression model to predict the likelihood of reengagement; at the second stage, we use Cox proportional hazards model to predict the proportional hazardous function of these reengaged customers. Since logistic regression model building process is mature in industry, we do not discuss this process in detail here.

BACKGROUND

To demonstrate the key ideas of this paper, we will walk through a case study of applying a two-stage model for a national wide multichannel retailer T, which sells a variety of products, including electronics, cosmetics, and shoes. In this case study, our goal is to win the most valuable inactive customers back by reactivation campaign. For Retailer T, inactive customers are defined as those whose last purchase date is between 12 months and 24 months before study start date. This time window is selected because those customers are suitable for a reengagement offer for Retailer T. For other businesses, such as furniture sales, diapers sales, the time window for inactive customers may vary (larger for furniture sales and smaller for diapers sales).

In this paper, we first focus on predicting the propensity of customers to respond to the campaign. Logistic regression model can be utilized for this purpose. Furthermore, our marketing goal is to reengage the most valuable customers, who will engage longer with Retailer T. Therefore, by applying survival model in the second stage, the reengagement rate is complemented by longevity. As Retailer T sends out millions of promotions each year, any significant lift in higher value customers will turn to significant lift in revenue.

DATA SOURCE

In this case, only inactive customers have been selected as our modeling base. The dataset has 50 masked transactional or demographical input variables with about one million customers. These 50 masked input variables are most predictive for the output responses. Variable selection is based on these 50 variables. The dataset has three desired response variables - Ind_resp, Resp_trips and Longevity.

- Ind_resp is a binary indicator - whether a customer purchases or not in Retailer T within one year after study start date (if yes, Ind_resp=1; else Ind_resp=0).
- Resp_trips is defined as how many trips a customer makes in Retailer T within one year after study start date.
- Longevity is defined as the days between first purchase date and last purchase date for a customer after reengagement, which can be interpreted as a customer's active lifetime in Retailer T within one year after the study start date. If a customer has never purchased, its longevity equals to 0.

STAGE I: LOGISTIC REGRESSION MODEL

At the first stage of our modeling, we apply logistic regression model to predict the likelihood to repurchase for inactive customers. Ind_resp is the response variable in our logistic regression model. In the selected customer base here, the reengagement rate is around 20%, which means that 20% of the inactive customers will purchase at least once in Retailer T in the next 12 months. Traditionally, logistic regression model is used to select customers with higher predicted reengagement probability.

On the other hand, from a business perspective, for an effective marketing campaign, there are two assumptions implied here:

1. The campaign can increase the reengagement rate of customers;
2. Customers with higher reengagement probability are more valuable and thus worth to invest on.

The first assumption is generally true. For the second assumption, we assume the customers who are more likely to respond are also the most valuable customers. However, although response likelihood and customer value are highly correlated, they are not exactly same.

Next, we will demonstrate our key focus: how to exploit survival analysis to select the most valuable customers for the reactivation campaign. Survival analysis is a good way to model the true long-time value of a customer.

STAGE II: SURVIVAL ANALYSIS

Exploratory Data Analysis

Before survival modeling, the distribution of longevity is profiled using PROC UNIVARIATE. Figure 2 shows the cumulative distribution function of longevity in the next year for the inactive customers. Customers who have never purchased or purchased only once are excluded in Figure 2. It gives us a raw picture of how those customers behave after reactivation. The smoothly increasing curve indicates a relatively even distribution in longevity for those reengaged customers.

Figure 2 is produced using the SAS code below:

```
PROC UNIVARIATE DATA=&data_in.(WHERE=(resp_trips>1));
    VAR longevity;
    CDFPLOT longevity;
RUN;
```

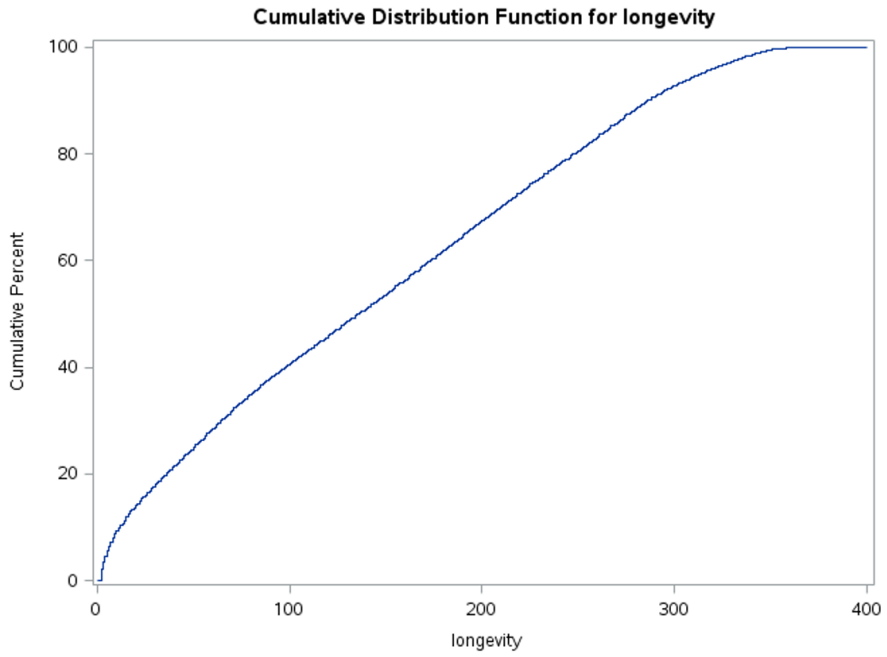

Figure 2. Cumulative Distribution Function for Longevity.

After the profiling of longevity, PROC LIFETEST is utilized to examine the pattern of survival probability on longevity by creating the Kaplan-Meier estimator of the survival function (see Figure 3) using the code below:

```
PROC LIFETEST DATA=&data_in. PLOTS=survival(cb) OUTS=&data_out. NOTABLE;
     TIME longevity*resp_trips(0,1);
RUN;
```

Because of the large dataset, we suppress the output table by specifying NOTABLE option in PROC LIFETEST. Non-reactivated customers and one-time buyers are censored out with longevity*resp_trips(0,1). With the option PLOTS=survival(cb), the confidence bands are added into Figure 3.

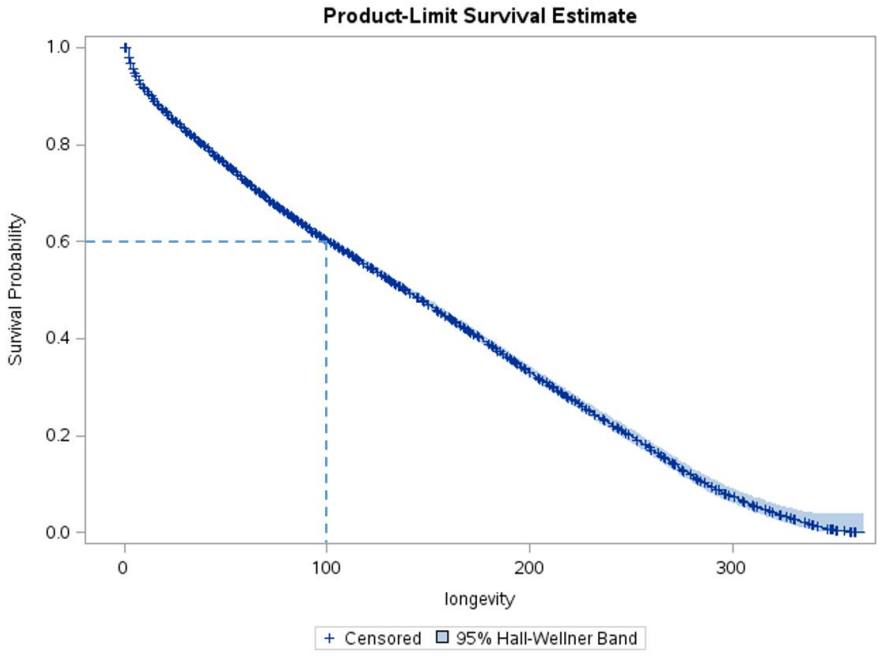

Figure 3. Kaplan-Meier estimator.

Here, Kaplan-Meier estimator can be interpreted as a measure of the fraction of customers still engaged with retailer T for a certain amount of time after reengagement. For instance, in Figure 3, 60% customers remain engaged for a longevity of 100 days. The smoothly dropping Kaplan-Meier curve indicates a stable disengaging rate for reengaged customers.

Cox Proportional Hazards Regression

After the data exploration, PROC PHREG is used for Cox proportional hazards regression modeling. To remove collinearity, variable selection is executed within the 50 predictors for retailer T. In the PHREG procedure, there are 5 variable selection methods available, which can be specified with SELECTION= in the model statement:

- NONE: fits the complete model specified in model statement
- FORWARD: forward selection
- BACKWARD: backward selection
- STEPWISE: stepwise selection
- SCORE: best subsets selection

The variable selection can be coded in SAS:

```
PROC PHREG DATA=&train_data.;
   ID acct_id;
   MODEL longevity*resp_trips(0,1)=&all_vars. / SELECTION=score BEST=3
   START=6 STOP=10;
```

```
RUN;
```

Here, SELECTION=score uses the branch-and-bound algorithm of Furnival and Wilson (Ref 1) to find the specified number of models with the highest likelihood (chi-square) for specified model sizes. For example, START=6 specified the minimum model size 6, STOP=10 specified the maximum model size 10, and BEST=3 specified the output to show the best 3 models for each model size from 6 to 10. The ID statement specifies the variable for identifying observations in the input data:

```
PROC CORR DATA=&train_data. PEARSON;
    VAR &surv_vars.;
RUN;
```

PROC CORR is used to check the correlation between each pair of variables. We specified Pearson product-moment correlation with PEARSON in CORR statement. The variable selection and correlation checks are repeated until high collinearity among variables are removed. 8 out of 50 input variables are selected for the final model. With the following procedure, a Cox proportional hazardous model is fitted:

```
PROC PHREG DATA=&train_data. PLOTS=survival;
    ID acct_id;
    MODEL longevity*resp_trips(0,1)=&surv_vars.;
    STORE store.survival_model;
RUN;
```

PLOTS=survival is specified to generate a reference curve of survival function (Figure 4) at the reference level of all categorical predictors and at the mean of all continuous predictors, which is the baseline hazardous function. The proportional hazardous function is fitted by the regression on all input predictors. The context and results of the statistical results are stored in store.survival_model that can be processed by PROC PLM. In library store, survival_model is saved as an item store file (in binary file format, Ref 2). Although it's hidden from view in SAS Explorer window, it's still a member of SAS library.

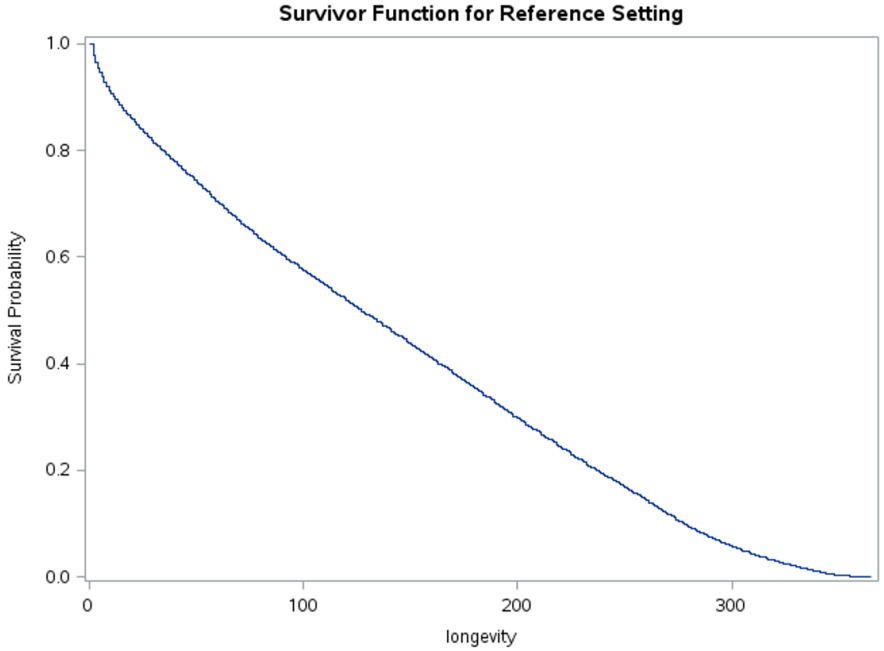

Figure 4. Survivor Function for Reference Setting.

The scoring can be coded in SAS:

```
PROC PLM RESTORE=store.survival_model;
    SHOW parms cov;
```

```
    SCORE DATA=&valid_data. OUT=valid_scored_out;
RUN;

DATA valid_scored_out;
    SET valid_scored_out;
    part_hazd=exp(predicted);
RUN;
```

Here, the fitted model was used by RESTORE= option in PROC PLM statement. The SHOW statement requests the parameter estimates and the covariance matrix of the parameter estimates in the fitted model. With the DATA step to convert the predicted value to partial hazardous prediction, we can then use PROC RANK to rank customers based on the partial hazardous prediction as they have the same baseline hazardous function:

```
PROC RANK DATA=valid_scored_out GROUPS=10 OUT=valid_scored_out;
    VAR part_hazd;
    RANKS surv_rank;
RUN;
```

In this way, lower ranks are given to customers with lower hazardous prediction, i.e. greater predicted longevity. Then, further marketing strategies can be applied to maximize revenue.

RESULTS

In our case study, our goal is to select the most valuable customers. Typically, in a marketing campaign, customers with high predicted value will be selected for promotion. Next, how to use the two-stage model to optimize marketing strategy is demonstrated.

The validation dataset is scored by the fitted logistic regression model and survival model. The scoring generate two predictions, which are likelihood of reengagement and proportional hazardous prediction. Then, logistic regression rank and survival model rank are created based on the two predicted values, respectively. Table 1 is the cross table of customer distribution in percentage on logistic regression rank by survival model rank for our validation dataset. From the table, generally, logistic regression rank and survival model rank are correlated as most customers are distributed along the diagonal line. Our campaign selection is based on Table 1.

Percentage		Survival Model Rank										Total
		0	1	2	3	4	5	6	7	8	9	
Logistic Regression Rank	0	6.68%	2.64%	0.53%	0.06%	0.01%	0.01%	0.01%	0.02%	0.02%	0.01%	10.00%
	1	1.92%	3.52%	3.03%	1.10%	0.22%	0.05%	0.02%	0.02%	0.03%	0.10%	10.00%
	2	0.86%	1.97%	2.52%	2.81%	1.23%	0.29%	0.06%	0.02%	0.03%	0.21%	10.00%
	3	0.39%	1.16%	1.89%	2.18%	1.97%	1.49%	0.39%	0.10%	0.04%	0.40%	10.00%
	4	0.11%	0.50%	1.21%	1.65%	1.72%	1.93%	1.68%	0.51%	0.15%	0.53%	10.00%
	5	0.02%	0.14%	0.51%	1.21%	2.45%	1.64%	1.51%	1.51%	0.45%	0.56%	10.00%
	6	0.01%	0.05%	0.23%	0.65%	1.47%	2.29%	1.62%	1.39%	1.50%	0.81%	10.00%
	7	0.00%	0.01%	0.08%	0.25%	0.71%	1.56%	2.72%	2.24%	1.21%	1.20%	10.00%
	8	0.00%	0.00%	0.02%	0.08%	0.19%	0.62%	1.57%	2.83%	2.88%	1.81%	10.00%
	9	0.00%	0.00%	0.00%	0.01%	0.03%	0.12%	0.42%	1.36%	3.70%	4.37%	10.00%
Total		10.00%	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%	10.00%

Table 1. Customer distribution by Logistic Regression Rank and Survival Model Rank (promotable customers by two-stage model are in the green filled region).

To check the performance of the two-stage model, for each cell in Table 1, the average longevity of customers is calculated in Table 2. From the table, generally, from top to bottom (by logistic regression rank) or from left to right (survival model rank), the average longevity has a decreasing trend. On some fields with few customers (match back to Table 1), the average longevity is not stable, thus of no significant impact.

Longevity		Survival Model Rank										Total
		0	1	2	3	4	5	6	7	8	9	
Logistic Regression Rank	0	51.2	38.0	29.5	26.5	47.9	22.2	17.7	18.2	22.0	16.7	46.2
	1	36.8	29.9	25.6	19.9	16.2	17.3	19.4	18.9	18.8	16.8	28.3
	2	25.8	29.3	22.8	18.0	13.9	16.2	21.3	17.5	14.7	12.5	21.4
	3	19.8	25.5	23.4	19.5	16.1	14.5	13.6	14.6	11.4	10.7	18.9
	4	17.0	20.9	20.2	18.2	16.2	16.6	14.3	13.6	11.5	8.7	16.4
	5	12.0	15.7	16.6	13.8	9.1	14.0	15.8	13.9	11.4	7.5	12.7
	6	18.0	11.5	13.1	13.9	10.2	10.6	14.1	13.2	11.1	7.6	11.6
	7	0.0	12.4	8.4	9.4	9.2	9.5	9.9	11.5	11.0	7.0	9.9
	8	0.0	15.2	9.9	6.4	8.4	7.9	6.7	7.2	7.8	5.5	7.0
	9	NA	0.0	0.1	8.8	4.7	4.1	4.8	4.8	4.5	3.5	4.1
Total		44.5	30.7	23.1	17.5	12.6	12.7	11.7	10.1	7.8	5.7	17.6

Table 2. Longevity Distribution by Logistic Regression Rank and Survival Model Rank.

To take the advantage of the two-stage modeling, for reactivation campaign, we can select those customers in the upper-left corner with relatively higher logistic regression rank and higher survival model rank, i.e. the green filled region in Table 1. In this way, the predicted reengagement rate is complemented with the predicted longevity of customers. The strategy optimizes the campaign selection, in which way, for retailer T, the output revenue is increased. In our study, compared with random selection, the selected top 30% customers using logistic regression rank have 164% lift in sales (Figure 5). On the other hand, the lift in sales of selected top 30% customers by two-stage model is 169% (Figure 5) compared with random selection, which is 5% higher than those customers selected using the logistic regression rank only. Different validation datasets are used to check the performance of this two-stage model and the results are stable.

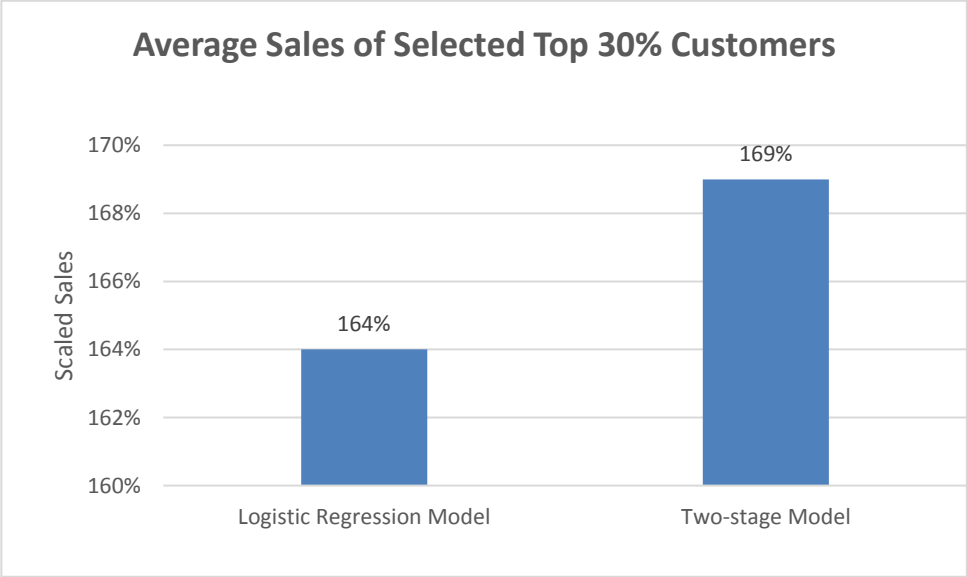

Figure 5 Average scaled sales of top 30% customers selected by different models. Sales are scaled by average sales of all customers.

CONCLUSION

By implementing a two-stage model, we are able to predict not only a customer's propensity to reengage, but also how long a customer will stay engaged. By layering survival model into traditional logistic

regression model, the likelihood of engagement is complemented by the long-time value of customers, thus the revenue of the reengagement campaign will be increased within our case study for retailer T.

REFERENCES

1. George M. Furnival and Robert W. Wilson, “Regression by Leaps and Bounds.”, Technometrics, Vol. 16, No. 4, Page 499–511 (1974).
2. Randall Cates, “What’s in a Name: Describing SAS File Types”, SUGI 27 Proceedings, Page 69 (2002).

ACKNOWLEDGMENTS

We would like to thank Ning Ma, Yin Chen and Delali Agbenyegah for encouraging us to write this paper and reviewing this paper. We also would like to thank the Predictive Analytics team within Marketing Insights at Alliance Data Card Services for their support.

CONTACT INFORMATION

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

Chao Xu
Enterprise: Alliance Data Systems
Address: 3100 Easton Square Pl, Columbus, Ohio 43219
E-mail: Chao.Xu@alliancedata.com

Jing Ren
Enterprise: Alliance Data Systems
Address: 3100 Easton Square Pl, Columbus, Ohio 43219
E-mail: Jing.Ren@alliancedata.com

Hongying Yang
Enterprise: Alliance Data Systems
Address: 3100 Easton Square Pl, Columbus, Ohio 43219
E-mail: Hongying.Yang@alliancedata.com