Objective

Orance Mobile is planning to implement a Pro-active Customer Retention Programme. In a nutshell, the programme aims to identify potential churned customers ahead of time and offer them with the appropriate incentive to entice them to stay. The objective of this project are as follows:

1. To develop predictive models for forecasting when a customer will churn.
2. To use these newly gained understanding for formulating appropriate customer retention incentive for customers who are high risk of churning 3 months after cutoff date for the analysis

Introduction

In this project, we will help Orange model which is mobile service operator with approximately 1 million subscribers to build the predictive models to forecast when a customer will churn and provide customer retention strategy to customer who are in the high risk of churn within 3 months after cut-off date.

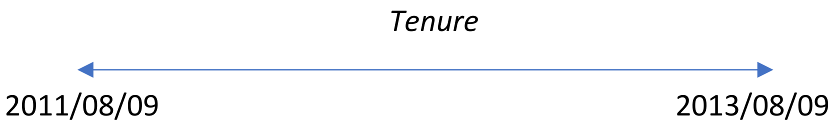
1 Data preparation

Let look at the original dataset. This dataset consists of 479357 observations and 23 variables, and each observation represents 1 unique customer.

1.1 Mobile dataset

Tenure and cut off date

We look at the variable STARTDATE, the start date of the subscription varies from 1996/11/01 to 2003/08/09 and the end date of the subscription varies from 1999/01/24 to 2013/08/09. As our analysis is only on 2 years period basis and the cut off date is 9th August 2003, our monitoring period is from 9th August 2001 to 9th August 2003 and we will remove the customers whose start date is before 9th August 2001 and customer whose stop date is after 9th August 2003.



Cause-specific and non-cause-specific event

We make a comparison between 2 variables STOPDATE and CONTRACTENDDATE. The stop date can be earlier, equal or later than the contract end date. We classify the churn cases into two types. One is churn due to end-of-contract period and another one is churn not due to end-of-contract period. We create the variable called target, if is account is still active, we will put 0 for target. If the account is inactive and stop date is between contract end date minus 30 days and contract end date plus 30 days, we will mark it as churn due to end-of-contract period. If the account is inactive and stop date is less than contract end date minus 30 days or greater than contract end date plus 30 days, we will mark it as churn not due to end-of-contract period.

Values	Explaination	Condition
Target = 0	Account is still active.	ISACTIVE = 1
Target = 1	Customer churn due to end-of-contract period.	ISACTIVE = 0 CONTRACTENDDATE – 30 days ≤ STOPDATE ≤ CONTRACTENDDATE + 30 days
Target = 2	Customer churn not due to end-of-contract period.	ISACTIVE = 0 STOPDATE < CONTRACTENDDATE - 30 days OR STOPDATE > CONTRACTENDDATE+ 30 days OR Missing value in CONTRACTENDDATE

1.2 Mobile_score dataset

As we need to score our current customer based on the Time-to-Event analysis results and find the customer who have high risk in the 3 months after cut-off date, we should prepare the score dataset for mobile.

This dataset will only include the record of customer who is still active. Therefore, from the cleaned mobile dataset, we remove the records whose ISACTIVE = 0.

And we also add a tenure variable named _t_. This variable contains the number of months since activation for each record. It can be derived by calculate the number of months between STARTDATE and CUTOFFDATE as shown below.

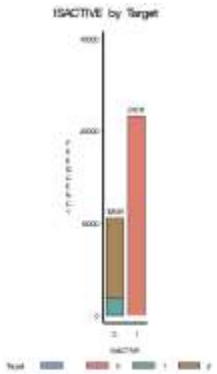
$$\_t\_ = month\ different(\ \mathrm{CUTOFFDATE} - \mathrm{STARTDATE}\ )$$

2 Data exploration and sampling

2.1 Data Exploration

We look at the time different between STARTDATE and CONTRACTENDDATE which is contract duration and notice that majority of contracts is 1-year contract.

We also take a look at account status at cut-off date 2013/08/09. In that point of time, 67.4% of accounts are still active, 5.7% of accounts are inactive due to end-of-contract and 26.9% of accounts are inactive not due to end-of-contract.



2.2 Data Partitioning

We partitioned this mobile dataset into 60% training data and 40% validation data.

Training data	Validation data
60%	40%

3 Time-to-Event analysis

We apply the Kaplan-Meier approach for survival analysis and try the model without any covariates and model with selected covariates.

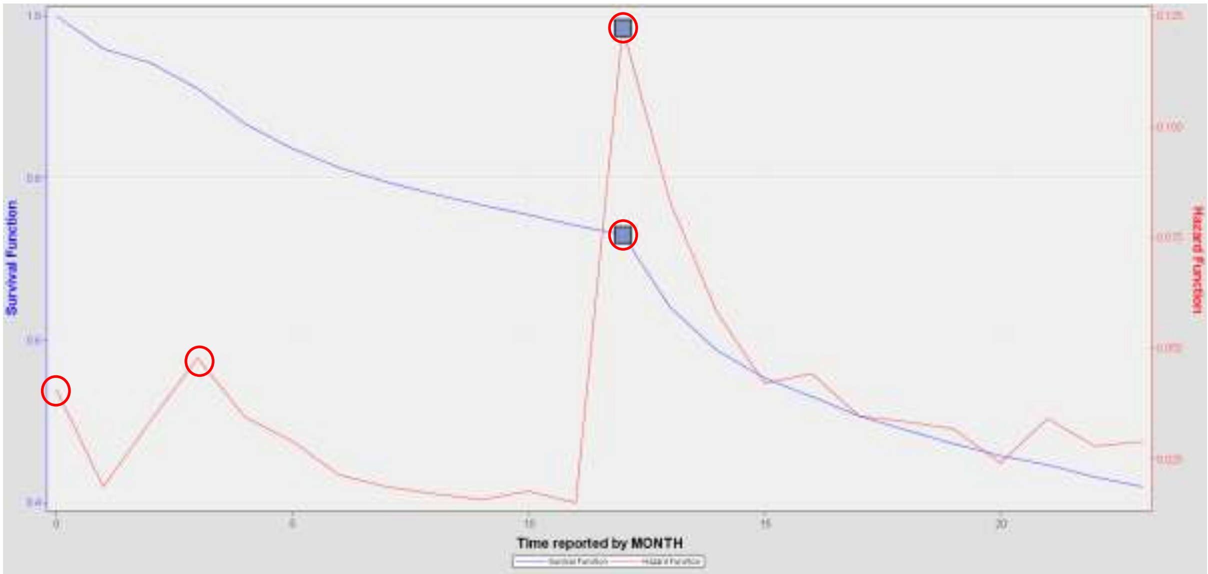
In the first model, I didn't add any covariates, only keep the 2 TimeID, 1 ID and 1 target.

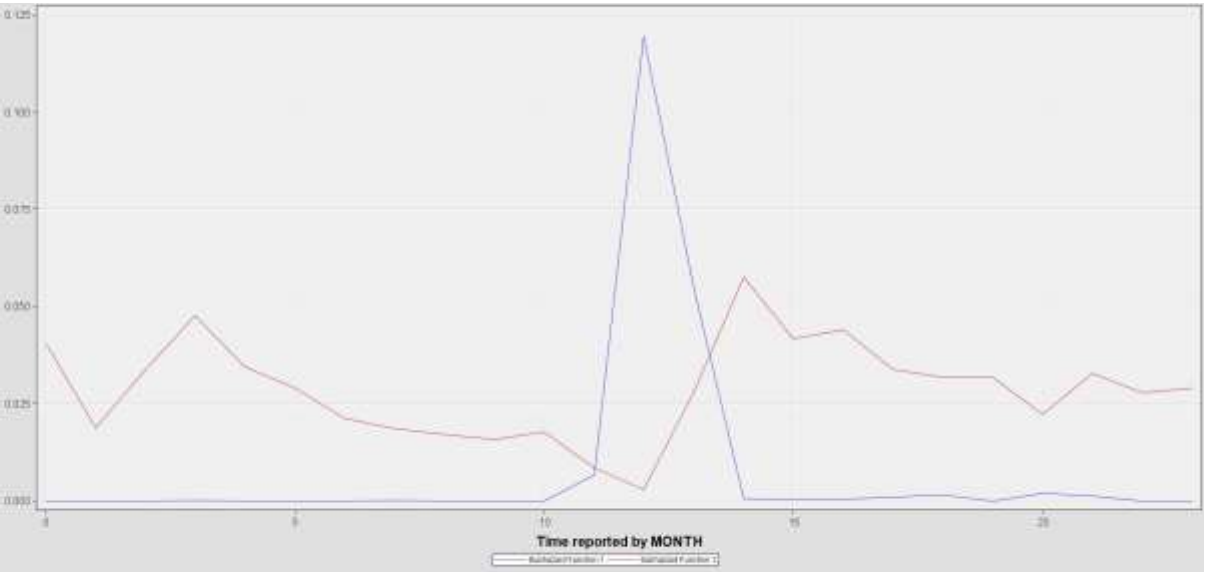
In the second model, I drop two variables ACCSTATUS and ACCSTATUSDTL, then use variable selection node to filter the variables with low R square value to the target. After variable selection, only 8 covariates were left, they are ACCTYPE, AGE, CONTRACTDURATION, CREDIT, autopay, DEPOSIT, MAKRER, NPA and INITMONTHLYFEE.

We compared two model's model validation graph and statistics and find that the model with covariate variables have better performance in terms of benefit value and concentration curve.

3.1 Empirical Survival Function for Training data

The empirical survival function empirical sub-hazard function for training data is shown in the figure below. In empirical survival function, the blue line is the survival function while the red line is the hazard function over 2 years period. And in empirical sub-hazard function, the blue line is the event of churn due to end-of-contract period while the red line is the event of churn not due to end-of-contract period.





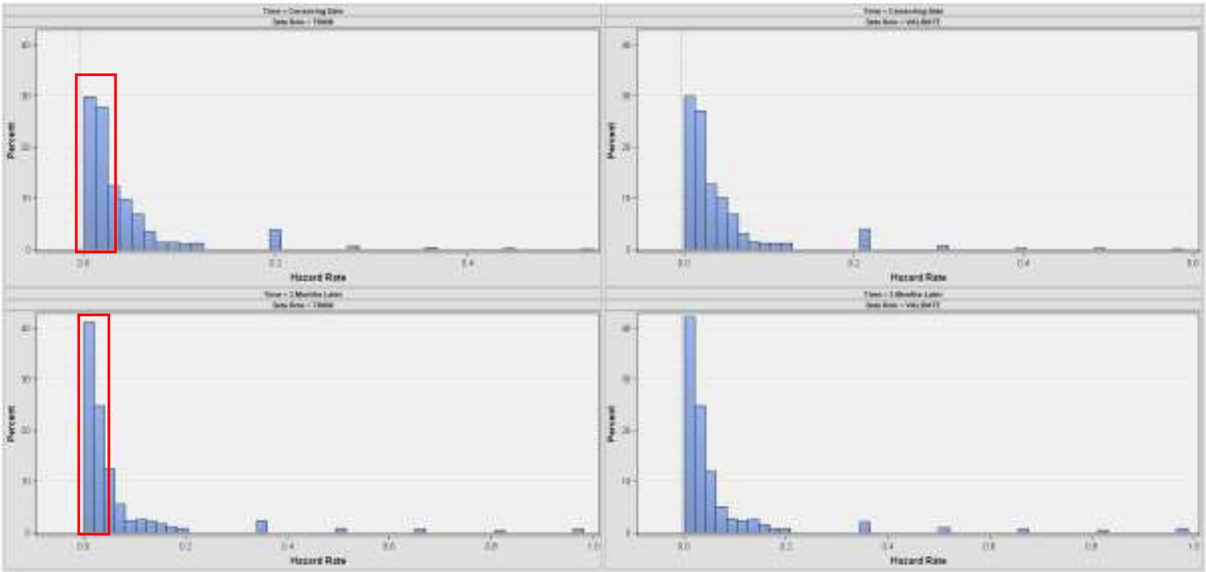
From above figures, we noticed that after the tenure of 24 months, about 42% of the customers are still active and more than half of the customers are left during that time period. There is one highest peak of hazard probability appeared at 12months after the start date. From the 11th month to 12th month, the hazard probability suddenly raised significantly from 1.52% to 12.24%, which means in 11th month, only 1.52% of customers churn but in 12th month, the percentage of customer churn case is 10 times higher than the previous month and raised to 12.24%. it seems logical for mobile operator that as majority of contracts is 1-year contract, the peak is corresponding to the ending of contract period in which customers start to compare the service and price with other mobile plan or mobile operators and consider terminating their current plan.

Beside the highest peak of hazard function, there is another peak appeared at 3rd month, the hazard probability raised to 4.78%. This maybe correspond to the ending of promotion period which is very common in cellphone industry that customer would like to terminate the service right after the promotion expires.

Last but not least, in time zero, the hazard probability is 4.05%. This reveals that there is a probability of 4.05% which customer will terminate the service right away after they subscribed. This is so called buyer's remorse which the sense of regret after making the purchase.

3.2 Hazard rate histogram

The hazard rate histogram of training data and validation data for the censoring date and 3 months after censoring date is shown in the figure below.

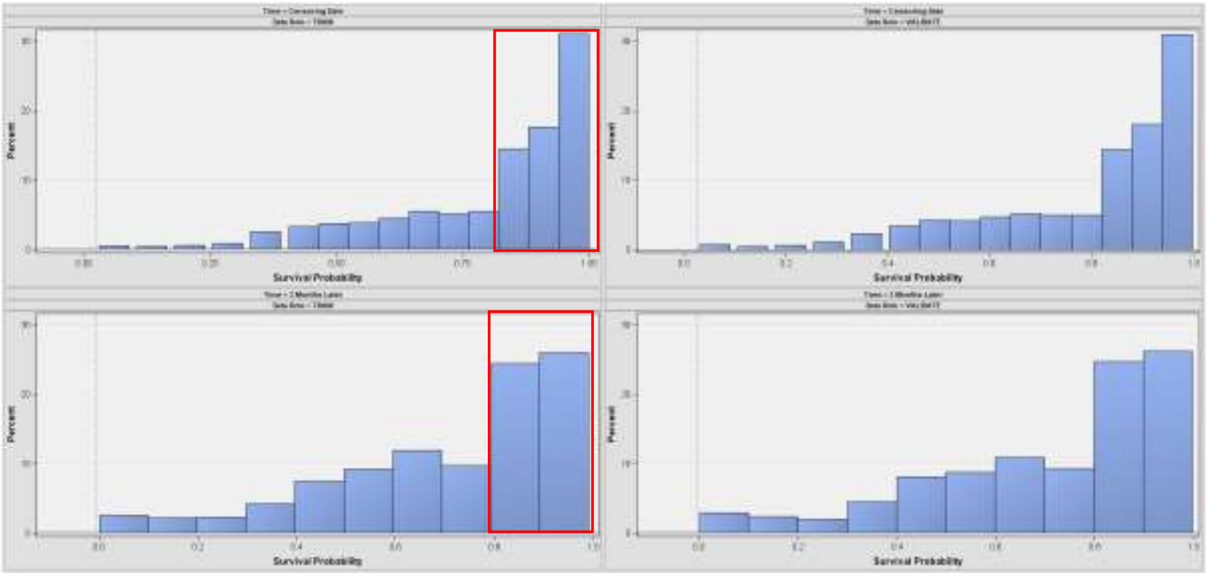


From the histogram in the top-left corner, we noticed that after 2 years from the starting date, 58% of customer in training dataset have a hazard rate between 0% to 2.5%, which means that 58% of customers have up to 2.5% probability to terminate their subscription, which is quite good result.

Furthermore, from the histogram in the bottom-left corner, we noticed that after 2 years and 3 months from the starting date, 66% of customers have up to 4.1% probability to terminate their subscription. Therefore we need to take attention to the customers who have high chance of terminate the subscription.

3.3 Survival probability histogram

The survival probability histogram of training data and validation data for the censoring date and 3 months after censoring date is shown in the figure below.

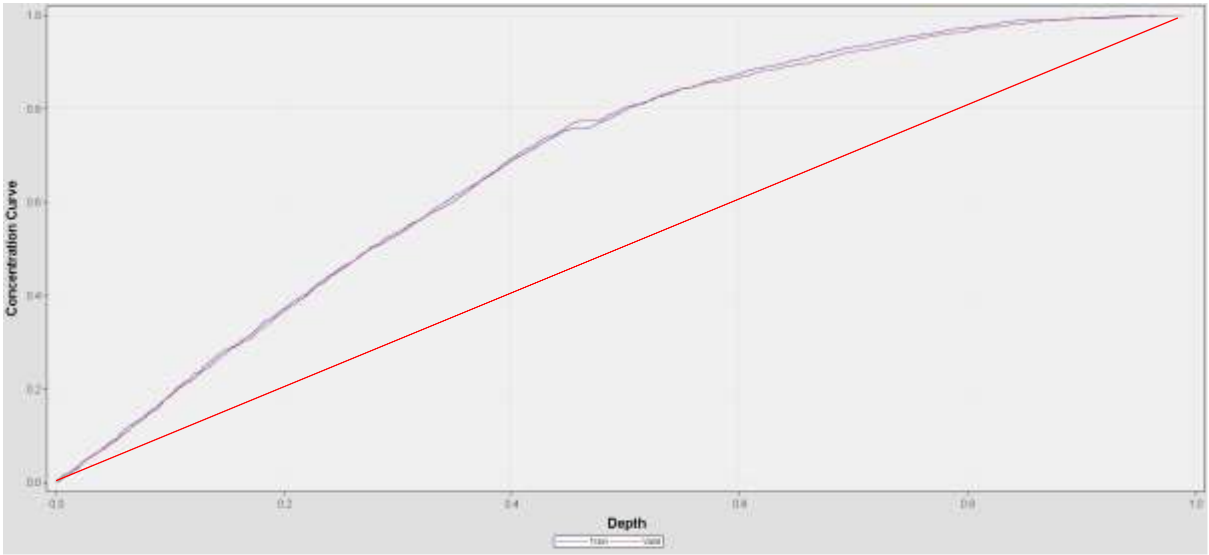


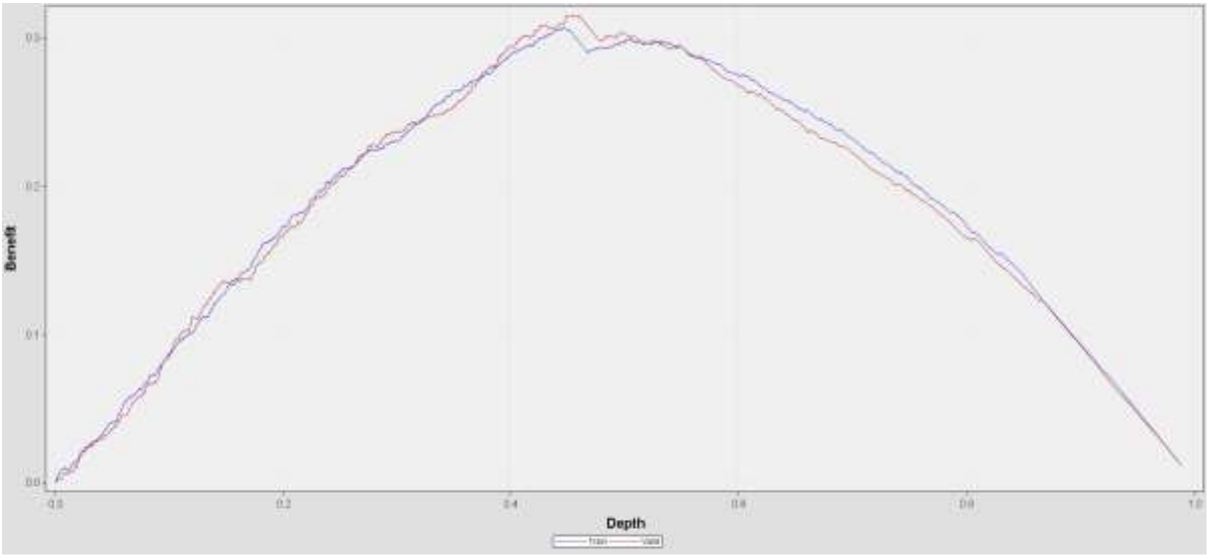
From the histogram in up-left corner, we noticed that 63.4% of customers have a survival probability between at least 81.53%. It reveals that after 2 years from starting date, there are 63.4% of customers have a high probability which is up to 81.53%to still subscribe our service.

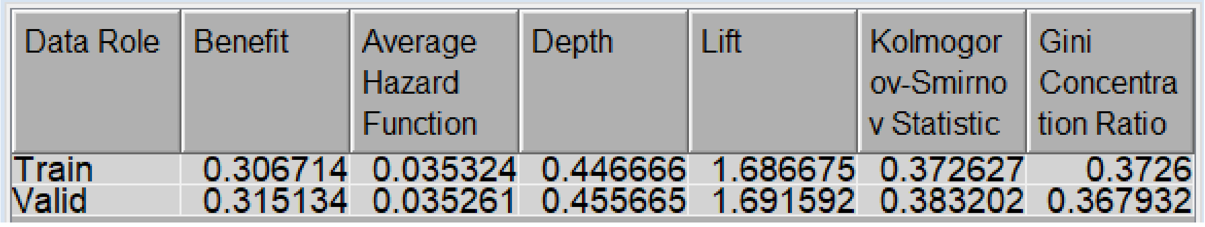
Furthermore, from the figure in bottom-left corner, we noticed that after 2 years and 3 months from the starting date, there are 50.5% of customers have a probability of above 80% to still subscribe our service.

3.4 Model validation plot and statistics

The concentration curve, benefit curve and statistics are shown as below. We add a diagonal line to compare our model’s performance with the random case.







Data Role	Benefit	Average Hazard Function	Depth	Lift	Kolmogorov-Smirnov Statistic	Gini Concentration Ratio
Train	0.306714	0.035324	0.446666	1.686675	0.372627	0.3726
Valid	0.315134	0.035261	0.455665	1.691592	0.383202	0.367932

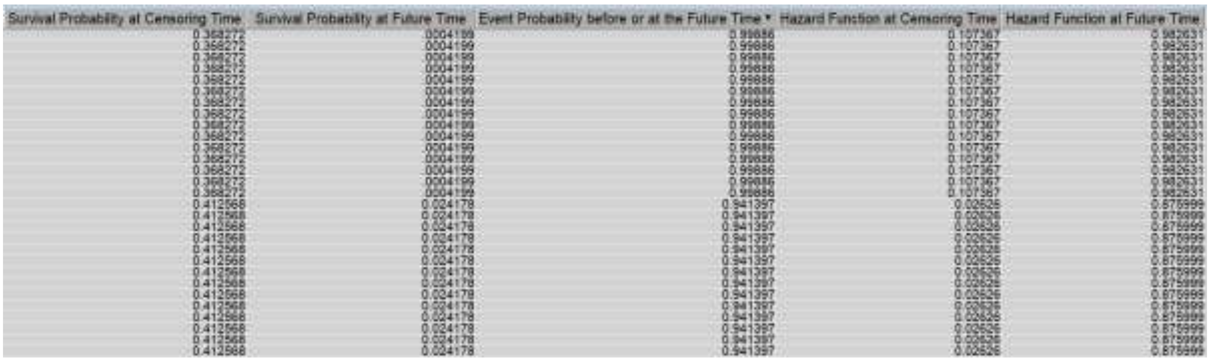
From the above concentration curve, we noticed that the concentration curve is above and have distance with the flat concentration which means the performance is not bad.

Moreover, in the above benefit curve of validation dataset, when the predicted probabilities at 0.6101, this model can best differentiate between the customers who will churn and customer who will not, and the best benefit at this point is 0.307

Last but not least, the difference of performance between train data and validation data is quite small which indicate a stable model and we are not overfitting.

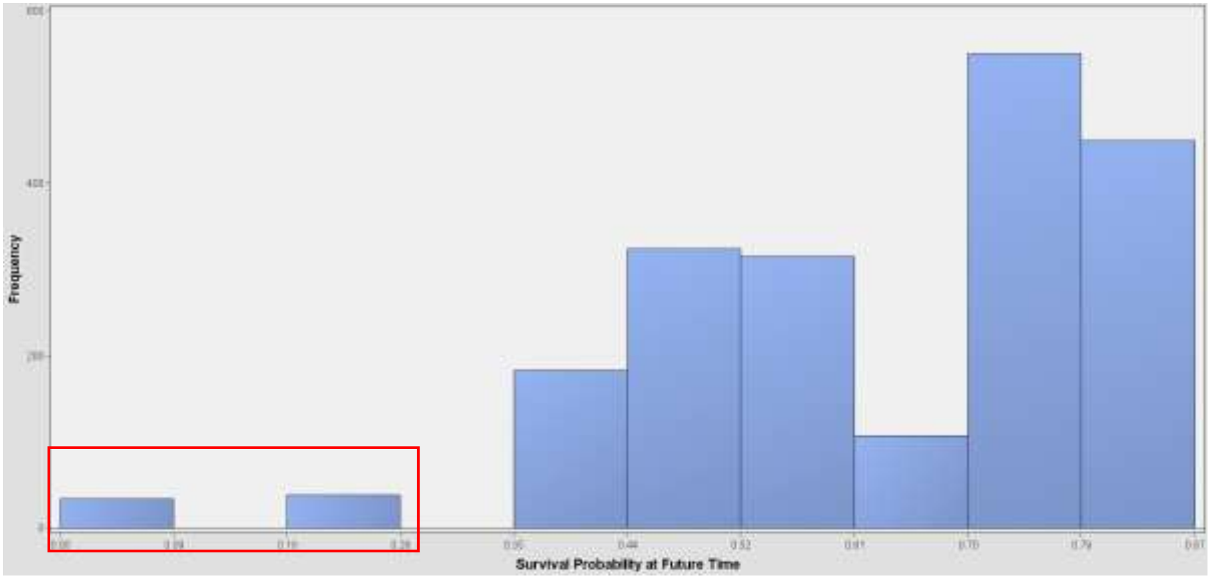
3.5 Score existing customer

Now we score the mobile_score data with the survival model. There are 4 important newly added columns survival probability at censoring time, survival probability at future time, event probability before or at the future time, hazard function at censoring time and hazard function at future time.



Survival Probability at Censoring Time	Survival Probability at Future Time	Event Probability before or at the Future Time ▾	Hazard Function at Censoring Time	Hazard Function at Future Time
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.368272	0004199	0.99886	0.107367	0.982631
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999
0.412568	0.024178	0.941397	0.02626	0.875999

Furthermore, let’s have a look at the survival probability distribution after 3 months.



From above distribution figure, we find that there are a few customers who have an extremely low survival probability which is up to 9% after 3 months, we also have almost half of the customers who have a survival probability greater than 70%. And the average survival probability is 62%.

In our next customer retention strategy, we focus on customers whose survival probability is less than 27%, they have a very high risk of terminate the service.

We should provide different kinds of promotion due to different reason of subscription. So we look at the starting date of each customers, if the starting date is within one month from the current date, we treat them as the customer who churn due to promotion period end, and we will give them another promotion for next three month but with lower discount than previous promotion. For the customers who churn due to contract end, we will give some promotion on recontract. For the customers who stay longer than contract end date, we treat them as our loyal customer, we can give more intensive or discount for them in order to keep these valuable customers.