Logistic Regression for Lapse Analysis

Tam Pham

4/8/2021

# Prevent Policies Lapse Proactively

Persistency is a key driver for successful insurance business. We just cannot let existing customers churn and then terminate their policies. Data science proactive alerts and actions are necessary to address policy lapse.

The typical solution approach is to devise a logistic regression model to predict the likelihood of lapse of policies. There are several data points that go in as inputs to this model like the following:

* Customer Demographics – Gender, Age, Race, Income, Nationality, Marital Status
* Customer Interaction mode and frequency with company – Email, Phone, others (fax, letters)
* Number and type of insurance products customers have bought from the company
* Policy details – Agent, Sum Insured, Premium, Term
* Each event for the policy – Inception, Lapse, Claim, Reinstatement, Cancel, Surrender, Mature

The model output helps in predicting whether a certain customer profile is likely to lapse or not. It also provides indicators on the significant factors impacting lapse, for example, Age, Premium level ,Channel of customer, Customer interaction etc that can help you take focused actions.

# Logistic Regression Analysis

Logistic Regression Analysis is a statistical analysis technique to :

* **model** the probability of an event occurring depending on the values of the independent variables
* **estimate** the probability that an event occurs for a randomly selected observation versus the the probability that the event does not occur
* **predict** the effect of a series of variables on a binary response variable
* **classify** observations by estimating the probability that an observation in a particular category (such as Lapse or No-Lapse in our study)

## Basic review of Logistic Regression

In Logistic Regresion, we study the probability of event which has two value 0 and 1: P(y|x) where y=[0,1] and x as input variables. Logistic Regression models the relationship between the probability of response (y) and predictor (x) using a logistic function:

(1)

By algebraic equation, the logistic regression can be written in terms of Odds ratio (“OR”) - of the event y (in this study ‘y’ is the ***Lapse event***)

(2)

Odds ratios range from 0 to positive infinity:  
- A value greater than 1 equates to probability of event greater than 0.5 (or 50%).  
- A value less than 1 equates to probability of event less than 0.5

Taking natural log of both sides, we have log-odds (logit) of the probability

(3)

Logistic regression transforms as a linear regression to find the log-odds of the probability. By analyzing linear regression function in formula (3) we can find the intercept() and coefficient () in order to solve Logistic regression in formula (2) and (1) as well as find their Odds ratio and Probability .

With result, we can interpret: Increasing x by one unit changes

1. the log odds by , or
2. the OR(y) by , or
3. the P(y) by

## Basic assumption for Binary Logistic Regression

The logistic regression analysis have the following basic assumptions

1. Continuous independent variables have a linear relationship with response function.
2. Independent variables do not have multiple linear relationships. (Multicollinearity)

## Selection of independence variables

We normally evaluate Logistic Regression Model based on:

* “Goodness of fit” parameters
* Deviance
* Likelihood ratio test (LRT)
* Akaike Information Criterion (AIC)

In this study we will run analysis to define the best logistic model follow

* AIC based algorithm : Apply Stepwise (forward) algorithm or Backward algorithm to find model with lowest AIC.
* Bayesian Model Average (BMA) : Apply to find model with lowest BIC (Bayesian Information Criterion)

# Data Set

## For testing data

## Explanatory of Data Set

The sample data for analysis has totaling of 1340 policies (434 Lapse, 907 inforce). Data have 10 variables divide into 2 groups:

1. Six quantitative variables as : Age, Premium Amount, Number of Reinstated, Number of Claims, Number of Emails, Number of Calls.
2. Four qualitative variables as : Gender, Occupation Group, Coverage Period, Payment Term,

# Build & Select Logistic Regression Model

We will analyze Logistic Regression in two study models

* Study 1 : Analyze Lapse with Customer Demography and Policy Detail
* Study 2 : Analyze Lapse with Customer Event and Interaction

The Logit model in both study are likely :

## Study 1 : Analyze Lapse with Customer Demography and Policy Detail

P is the probability of Lapse based on the following factors : Sex, Occupation, Premium, Coverage Period, Payment Term. There are variation of models :

* Logit(P) = a + b\*Sex
* Logit(P) = a + b\*Sex + c\*Occupation
* Logit(P) = a + b\*Sex + c\*Occupation + d\*Premium
* Logit(P) = a + b\*Sex + c\*Occupation + d\*Premium + e\*Coverage
* Logit(P) = a + b\*Sex + c\*Occupation + d\*Premium + e\*Coverage + f\*Payment

### Using backward stepwise method to examine the selection of independent variables

## Start: AIC=1205.31  
## Lapse ~ PO\_Sex + Occupation + Premium + CoveragePeriod + PaymentTerm  
##   
## Df Deviance AIC  
## - PO\_Sex 1 1193.4 1203.4  
## - PaymentTerm 1 1194.9 1204.9  
## <none> 1193.3 1205.3  
## - Occupation 1 1198.1 1208.1  
## - CoveragePeriod 1 1219.5 1229.5  
## - Premium 1 1546.2 1556.2  
##   
## Step: AIC=1203.43  
## Lapse ~ Occupation + Premium + CoveragePeriod + PaymentTerm  
##   
## Df Deviance AIC  
## - PaymentTerm 1 1195.0 1203.0  
## <none> 1193.4 1203.4  
## - Occupation 1 1198.2 1206.2  
## - CoveragePeriod 1 1219.7 1227.7  
## - Premium 1 1546.4 1554.4  
##   
## Step: AIC=1202.99  
## Lapse ~ Occupation + Premium + CoveragePeriod  
##   
## Df Deviance AIC  
## <none> 1195.0 1203.0  
## - Occupation 1 1199.8 1205.8  
## - CoveragePeriod 1 1221.4 1227.4  
## - Premium 1 1547.2 1553.2

##   
## Call: glm(formula = Lapse ~ Occupation + Premium + CoveragePeriod,   
## family = binomial, data = dat1)  
##   
## Coefficients:  
## (Intercept) Occupation Premium CoveragePeriod   
## -3.1110538 -0.1868714 0.0001261 0.5100813   
##   
## Degrees of Freedom: 1340 Total (i.e. Null); 1337 Residual  
## Null Deviance: 1689   
## Residual Deviance: 1195 AIC: 1203

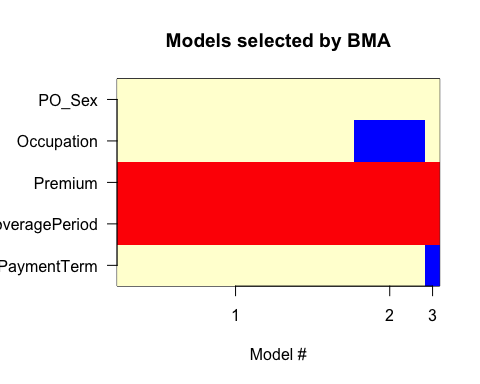
With smallest AIC, the suggested model for Logit(P) is:

Logit(P) = -3.1111+(-0.1869)\*Occupation + 0.00013\*Premium + 0.5101\*CoveragePeriod

*Sex and Payment Term are eliminated from suggested model.*

### Using Bayesian Model Average (BMA) to find model with smallest Bayesian Information Criterion (BIC)

##   
## Call:  
## bic.glm.data.frame(x = predictors, y = outcome, glm.family = "binomial", strict = F, OR = 20)  
##   
##   
## 3 models were selected  
## Best 3 models (cumulative posterior probability = 1 ):   
##   
## p!=0 EV SD model 1 model 2   
## Intercept 100 -3.4838178 3.478e-01 -3.602e+00 -3.111e+00  
## PO\_Sex 0.0 0.0000000 0.000e+00 . .   
## Occupation 22.2 -0.0414068 8.739e-02 . -1.869e-01  
## Premium 100.0 0.0001303 8.673e-06 1.315e-04 1.261e-04  
## CoveragePeriod 100.0 0.5050956 1.008e-01 5.037e-01 5.101e-01  
## PaymentTerm 4.4 -0.0036987 2.224e-02 . .   
##   
## nVar 2 3   
## BIC -8.435e+03 -8.433e+03  
## post prob 0.734 0.222   
## model 3   
## Intercept -3.389e+00  
## PO\_Sex .   
## Occupation .   
## Premium 1.316e-04  
## CoveragePeriod 5.035e-01  
## PaymentTerm -8.394e-02  
##   
## nVar 3   
## BIC -8.430e+03  
## post prob 0.044



BMA suggests best 3 models with cumulative posterior probability equal 1. Two of them have total post prob of 95%:

* Model 1 : Logit(P) = -3.602 + 0.00013\*Premium + 0.5037\*CoveragePeriod
* Model 2 : **Logit(P) = -3.111 - 0.1869\*Occupation + 0.00013\*Premium + 0.5101\*CoveragePeriod**

We consider Model 2 as it is also suggested via AIC algorithm, its Deviance is lower and LRT is statistical significance.

### Validation of logistic regression model

Validate the validity of selected model based on statistical value

## Logistic Regression Model  
##   
## lrm(formula = Lapse ~ Occupation + Premium + CoveragePeriod,   
## data = dat1)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 1341 LR chi2 493.55 R2 0.430 C 0.805   
## 0 907 d.f. 3 g 1.736 Dxy 0.610   
## 1 434 Pr(> chi2) <0.0001 gr 5.677 gamma 0.610   
## max |deriv| 2e-05 gp 0.286 tau-a 0.267   
## Brier 0.139   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -3.1111 0.3393 -9.17 <0.0001   
## Occupation -0.1869 0.0853 -2.19 0.0286   
## Premium 0.0001 0.0000 14.66 <0.0001   
## CoveragePeriod 0.5101 0.1011 5.05 <0.0001   
##

* Overall model evaluation by Likelihood ratio test had a p-value of 0.0001 which was less than significant level 0.05 , so the model statistically significant.
* Area under the curve (AUC) or c-value is to measure discrimination between observed and predicted value; for this model c-value is good at 0.805.
* Psuedo-R2 expresses the percentage of the variability of the outcome variable by the independent variables. From table below, the model account for 29.2%,30.8% or 43% based on statistic formula:

|  |  |
| --- | --- |
| Based on | Psuedo-R2 |
| Hosmer and Lemeshow | 0.292 |
| Cox and Snell | 0.308 |
| Nagelkerke | 0.430 |

## Study 2 : Analyze Lapse with Customer Event and Customer Interaction

We will investigate the Lapse possibility in correlation with Customer Events: Number of Reinstatement, Number of Claims and Interaction activities with customer such as Number of Emails, Phone calls.

### Using stepwise to identify lowest AIC model

## Start: AIC=1583.99  
## Lapse ~ NumOfReinstated + NumOfClaims + NumOfEmails + NumOfCalls  
##   
## Df Deviance AIC  
## <none> 1574.0 1584.0  
## - NumOfEmails 1 1586.9 1594.9  
## - NumOfClaims 1 1591.8 1599.8  
## - NumOfCalls 1 1604.3 1612.3  
## - NumOfReinstated 1 1669.5 1677.5

##   
## Call: glm(formula = Lapse ~ NumOfReinstated + NumOfClaims + NumOfEmails +   
## NumOfCalls, family = binomial, data = dat2)  
##   
## Coefficients:  
## (Intercept) NumOfReinstated NumOfClaims NumOfEmails   
## -0.6385 0.5673 0.2582 -0.2323   
## NumOfCalls   
## -0.3411   
##   
## Degrees of Freedom: 1340 Total (i.e. Null); 1336 Residual  
## Null Deviance: 1689   
## Residual Deviance: 1574 AIC: 1584

##   
## Call:  
## glm(formula = Lapse ~ NumOfReinstated + NumOfClaims + NumOfEmails +   
## NumOfCalls, family = binomial, data = dat2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6493 -0.8363 -0.7218 1.1149 2.4005   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.63851 0.10804 -5.910 3.43e-09 \*\*\*  
## NumOfReinstated 0.56733 0.06036 9.398 < 2e-16 \*\*\*  
## NumOfClaims 0.25823 0.06094 4.237 2.26e-05 \*\*\*  
## NumOfEmails -0.23230 0.06633 -3.502 0.000461 \*\*\*  
## NumOfCalls -0.34113 0.06474 -5.269 1.37e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1688.5 on 1340 degrees of freedom  
## Residual deviance: 1574.0 on 1336 degrees of freedom  
## AIC: 1584  
##   
## Number of Fisher Scoring iterations: 4

Logit(p) = -0.6385 +0.5673\*NumOfReinstated +0.2582\*NumOfClaims -0.2323\*NumOfEmails -0.3411\*NumOfCalls

### Using Bayesian Model Average (BMA)

##   
## Call:  
## bic.glm.data.frame(x = predictors2, y = outcome2, glm.family = "binomial", strict = F, OR = 20)  
##   
##   
## 2 models were selected  
## Best 2 models (cumulative posterior probability = 1 ):   
##   
## p!=0 EV SD model 1 model 2   
## Intercept 100 -0.6500 0.11752 -0.6385 -0.8527  
## NumOfReinstated 100.0 0.5643 0.06151 0.5673 0.5113  
## NumOfClaims 100.0 0.2570 0.06112 0.2582 0.2358  
## NumOfEmails 94.6 -0.2199 0.08305 -0.2323 .   
## NumOfCalls 100.0 -0.3424 0.06494 -0.3411 -0.3655  
##   
## nVar 4 3   
## BIC -8046.7741 -8041.0281  
## post prob 0.946 0.054

2 models are selected by BMA with cumulative posterior probability equal 1:

* Model 1 with post prob of 94.6% have coefficients as same as suggested in AIC algorithm:
* **Logit(P) = - 0.6385 + 0.5673\*NumOfReinstated + 0.2582\*NumberOfClaims - 0.2323\*NumOfEmails - 0.3411\*NumberOfCalls**
* Model 2 with 5.4% post prob :
* Logit(P) = - 0.8527 + 0.5113\*NumOfReinstated + 0.2358\*NumberOfClaims - 0.3665\*NumberOfCalls

Model 1 with BIC=-8046 is preferable than Model 2 with BIC=-8041

### Checking the validity of model

## Logistic Regression Model  
##   
## lrm(formula = Lapse ~ ., data = dat2)  
##   
## Model Likelihood Discrimination Rank Discrim.   
## Ratio Test Indexes Indexes   
## Obs 1341 LR chi2 114.55 R2 0.114 C 0.650   
## 0 907 d.f. 4 g 0.704 Dxy 0.300   
## 1 434 Pr(> chi2) <0.0001 gr 2.022 gamma 0.313   
## max |deriv| 2e-09 gp 0.146 tau-a 0.132   
## Brier 0.198   
##   
## Coef S.E. Wald Z Pr(>|Z|)  
## Intercept -0.6385 0.1080 -5.91 <0.0001   
## NumOfReinstated 0.5673 0.0604 9.40 <0.0001   
## NumOfClaims 0.2582 0.0609 4.24 <0.0001   
## NumOfEmails -0.2323 0.0663 -3.50 0.0005   
## NumOfCalls -0.3411 0.0647 -5.27 <0.0001   
##

## Pseudo R-squared for logistic regression model  
##   
## Hosmer and Lemeshow R-squared 0.068   
## Cox and Snell R-squared 0.082   
## Nagelkerke R-squared 0.114

* As p-value of Likelihood Ratio test less than significant value (<0.0001), the LR Test for model is statistically significant.
* C-value of 0.650 show a fairly discrimination between observed and predicted values.

# Interpretation of Coefficient and Forecasting

## Study 1: Lapse probability with Occupation, Premium and Coverage Period

**Logit(P) = -3.1110 - 0.1869\*Occupation + 0.0001\*Premium + 0.5101\*CoveragePeriod**

***Odds ratio table***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Odds ratios | CI 95% | for Diff. |
| Occupation | 0.829 | 0.702 - 0.981 | 1 |
| Premium | 4.362 | 3.583 - 5.312 | 11682 |
| CoveragePeriod | 2.773 | 1.866 - 4.122 | 2 |

Given others variables constant,

* For each unit increase of *Occupation group*, logit(P) decreases 0.187 units, Odds ratio (OR) of probability P increase 0.829 or *predict* Odds(P) decrease 17%.
* For each 1,000 Dollar increase of *Premium*, logit(P) decreases 1,000\*0.000126 units = 0,1261 units, OR(̂P) changes or *predict* Odds(P) increase 13.44%
* For each unit increase of *Coverage Period*, logit(P) increase 0.5101 units respectively , OR(P) changes or *predict* Odds(P) increase 66.55%

***Lapse Forecasting***

From the observed data, this predictive model forecasts 281 Lapse policies, representing of 20.95 %, and 1060 insured expecting do not terminate their policy (or ~79.05%); the accuracy of this forecasting is 0.8233 with 95% CI (0.8018, 0.8433) and Specificity=0.9537, Sensitivity=0.5507.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Observation** |  |  |
| **Forecasting** | **Policy Lapse** | **No-Lapse** | **Total** |
| Policy Lapse | 239 | 42 | 281 |
| No Lapse | 195 | 865 | 1060 |
| Total | **434** | **907** | **1341** |

## Study 2: Lapse probability with Number of Reinstated, Claims and Number of Emails, Calls to Customer

**Logit(P) = -0.6385 + 0.5673\*NumOfReinstated + 0.2582\*NumberOfClaims - 0.2323\*NumOfEmails - 0.3411\*NumberOfCalls**

***Lapse forecasting by model***

The predictive model in study 2 forecasting 146 Lapse case representing of 10.9% and 1195 No-Lapse case representing 89.1%; accuracy of this predicting model is 0.7271 with 95% CI(0.7071, 0,7508).

## actual\_values  
## predicted\_values 0 1  
## 0 868 327  
## 1 39 107

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 868 327  
## 1 39 107  
##   
## Accuracy : 0.7271   
## 95% CI : (0.7024, 0.7508)  
## No Information Rate : 0.6764   
## P-Value [Acc > NIR] : 3.23e-05   
##   
## Kappa : 0.2461   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.24654   
## Specificity : 0.95700   
## Pos Pred Value : 0.73288   
## Neg Pred Value : 0.72636   
## Prevalence : 0.32364   
## Detection Rate : 0.07979   
## Detection Prevalence : 0.10887   
## Balanced Accuracy : 0.60177   
##   
## 'Positive' Class : 1   
##

***Interpreting Odds and Probability***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Odds ratios | CI 95% | for Diff. |
| NumOfReinstated | 1.763 | 1.567 - 1.985 | 1 |
| NumOfClainms | 1.295 | 1.149 - 1.459 | 1 |
| NumOfEmails | 0.792 | 0.696 - 0.903 | 1 |
| NumOfCalls | 0.505 | 0.392 - 0.651 | 2 |

Given others variables constant,

* For each unit increase of *Reinstated*, Odds ratio (OR) of probability P increase 1.763 or *predicting* Odds(P) increase 76.3%.
* For each unit increase of *Claim*, Odds ratio (OR) of probability P increase 1.295 or *predicting* Odds(P) increase 29.5%.
* For each unit increase of *Email*, *predicting* Odds(P) decrease 20.7%
* For each unit increase of *Call*, *predicting* Odds(P) decrease 28.9%

## Study findings

From the observed information of insured, the predictive model in study 1 forecasting more accuracy with 82.23% (versus 72.21%) and more discriminated with C-value of 0.805. Among True Positive Lapse (predicted Lapse=observed Lapse), it was found that 212 Lapse policy (~88.7%) with Occupation in Group 1 and 2, 222 policies (~92.9%) has Coverage period in Group 2 and 3, and 89.5% lapse policies with premium over $2,000.

The model **Logit(P) = -3.1110 - 0.1869\*Occupation - 0.00013\*Premium + 0.5101\*CoveragePeriod** predict 281 lapse cases, accounting for 20.85%. Ability to accurate forecast of 82.33%.

The model **Logit(P) = -0.6385 + 0.5673\*NumOfReinstated + 0.2582\*NumberOfClaims - 0.2323\*NumOfEmails - 0.3411\*NumberOfCalls** predict 146 lapse cases, accounting for 10.9% total policies. This model show that increasing Number of Reinstatement and Claims can increase the possibility of Lapse; but by increasing interaction to customer, we expect to reducing the possibility of insured to terminate their policy.