Customer Churn Prediction

Python & SAS

Date: 5th April 2022
Created by: Varun Bhavnani
No. of pages: 8

Document Details

Document Name	Document Owner	Email ID	Date of release	Version
Customer Churn Prediction	Varun Bhavnani	bhavnani.varun.97@gmail.com	5th April 2022	V1.0

Contents

1. APPROACH 1 **4**

1. Estimating the best logit model 4

2. Which are the top three factors that affect churn in your model and what is their effect size? 5

3. What other variables (that if collected) would help to improve the fit of the model. 5

4. Compute the hit ratio for your model. Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data. 5

5. Using the model parameters predict the churn for the holdout sample as well and compute the hit ratio. 6

2. Approach 2: **7**

1. APPROACH 1

1. Estimating the best logit model

Interpretation:

- We calculated the percentage difference between mean values of the churners and the non-churners. We checked for the correlation between the variables. Based on the results, the 10 variables which we have selected are change_mou blck_dat_mean roam_mean drop_dat_mean mou_opkd_mean threeway_mean custcare_mean callfwdv_Mean plcd_dat_Mean callwait_Mean.
- On running the logistic regression we found roam_mean, threeway_Mean, custcare_Mean, callwait_Mean, and change_mou to be significant. However, we found that change_mou has no effect on Churn and hence we replaced change_mou with the next significant independent variable eqpdays in the percentage difference list.
- AIC & BIC/SC: From the table below, we can interpret that AIC & BIC/SC values for the 10 variables model were higher as compared to that of the reduced-form model with 5 variables. Also, Intercept & covariate columns are less than in the Intercept only column. Hence, our fitted model is better. The model is statistically significant.
- Significance of betas: All the variables have a p-value < 0.05 i.e all are statistically significant
- Prediction accuracy (Percent of Concordance): 57.5 percent of pairings in which the observation with the customer churning has a greater projected probability than the observation without the customer churning.
- Odds-ratio:
    - roam_Mean: For every 1 unit increase of Roaming, the odds of Customer Churn increase by 0.4%, holding everything else constant.
    - threeway_Mean: For every 1 unit increase of Three-Way Calls, the odds of Customer Churn decreased by 2.6%, holding everything else constant.
    - custcare_Mean: For every 1 unit increase of Customer Care Calls, the odds of Customer Churn decreased by 1%, holding everything else constant.
    - Eqpdays: For every 1 unit increase of the Number of days of the current equipment, the odds of Customer Churn increase by 0.08%, holding everything else constant.
    - callwait_Mean: For every 1 unit increase of callwait_Mean, the odds of Customer Churn decreased by 0.3%, holding everything else constant.

Table:

Variables	coefficients	t-Value	Odds Ratio
roam_Mean	0.00396	13.9282	1.004
threeway_Mean	-0.0261	10.7585	0.974
custcare_Mean	-0.0099	29.2809	0.99
callwait_Mean	-0.00289	3.306	0.997
eqpdays	0.000839	722.2832	1.001

AIC & BIC/SC:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	96694.897	95739.477
SC	96704.050	95794.393
-2 Log L	96692.897	95727.477

2. The top three factors that affect churn in your model and what is their effect size

- The top 3 models that affect customer churn are
  - roam_Mean: Effect size of 0.396%
  - threeway_Mean: Effect size of -2.61%
  - custcare_Mean: Effect size of -0.99%

3. Other variables (that if collected) would help to improve the fit of the model.

- Frequency: The frequency with which the customer uses tech support services, i.e. the number of encounters with the tech support department. The bigger the number of encounters, the greater the consumer involvement with the product or service. As a result, the likelihood of client attrition is reduced.
- Satisfaction ratings: The rating supplied by a customer after speaking with a customer service representative can be used to determine the customer's degree of satisfaction.
- The amount owed by the customer can be used to determine the likelihood of customer turnover. The more the amount owed, the greater the likelihood of customer turnover.

4. Computing the hit ratio for your model

- Using the logit model, we predicted the customers who are going to churn by considering. If prediction < 0.5 then churn =0, and if prediction > 0.5 then churn =1. We compared these predictions to the actual values and calculated the hit ratio which is 54.57%.

hit_ratio
54.57227

5. Using the model parameters, predicting the churn for the holdout sample as well and compute the hit ratio.

- Using the logit model, we predicted the customers who are going to churn by considering. If prediction < 0.5 then churn =0, and if prediction > 0.5 then churn =1. We compared these predictions to the actual values and calculated the hit ratio which is 41.52%.

hit_ratio
41.5061

2. Approach 2:

In this approach, we started with the same top 10 variables and then filtered them out based on which ones are statistically significant. The only difference in this approach is that instead of discarding those non-significant variables, we chose the next variables in our mean difference ranking and tested their significance. All we are doing here is maintaining the top 10 significant variables and discarding those which are correlated in the end. Following are the details regarding the new model we created.

Top 10 significant variables without correlation:

Churn = roam_mean threeway_mean custcare_mean eqpdays vceovr_Range hnd_price income totmrc_Mean

AIC & BIC/SC:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	71514.245	70266.234
SC	71523.096	70345.896
-2 Log L	71512.245	70248.234

We can see that our AIC & BIC are much better (lesser) than the one we got in our first approach

Table:

Variables	coefficients	t-Value	Odds Ratio
roam_Mean	0.00556	15.2713	1.006
threeway_Mean	-0.0405	15.9863	0.96
custcare_Mean	-0.0132	30.5472	0.987
eqpdays	0.000585	209.8798	1.001
vceovr_Range	0.00231	192.0967	1.002
hnd_price	-0.0023	183.548	0.998
income	-0.00905	4.7732	0.991
totmrc_Mean	-0.004	93.7426	0.996

Hit Ratio on Train data:

hit_ratio
42.0124

Hit ratio on Test data:

hit_ratio
41.5061

Percentage on Concordant:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.2	Somers' D	0.185
Percent Discordant	40.8	Gamma	0.185
Percent Tied	0.0	Tau-a	0.092
Pairs	665398742	c	0.592

Here, the percent concordant is improved, 59.2% as compared to what we got in our previous approach: 57.5%

This seems to be a better model for predicting the customer churn

- End of Document -