PREDICTING CUSTOMER CHURN AT QWE INC.

Peilin Zhong, Zheng Li

March 9, 2019

1 Executive Summary

This report provides an analysis of impact of customer features on customer churn. Our method of analysis are using single logistic regression and multiple logistic regression.

Our analysis shows that Customer Age does no have strong relationship with customer churn in the first place, and using single feature, CHIScoreMonth, is not accurate and sensible. But after using multiple logistic regression with CustomerAgeinmonths, CHIScoreMonth0, CHIScore01, SP01 and DaysSinceLastLogin01, we find out that Customer Age have impact on customer churn. And we conclude that DaysSinceLastLogin01, CustomerAgeinmonths and CHIScoreMonth0 have most impact on deciding customer churn.

2 Background

As a successful dot-com start-ups, after fast growth initially, QWE realized the need for deeper analytical insight into some key business processes, one of which was customer retention. At first, QWE tried to convince the customer to extend the contract by offering free services or discounts on existing services. However, QWE wondered if they could develop a more proactive approach. Also, they hoped they could estimate the probability that a given customer would leave in the near future and identify the drivers that contributed most to that customer's decision. To solve this problem, QWE wanted to generate a list of the 100 customers who were most likely to leave and, if possible, the three factors contributing most to that likelihood.

To collect dataset, QWE rolled back two months to December 1, 2011, and obtained a sample of 6,000 of QWE's customers as of that date. To start with this task, Customer age, CHI [Customer Happiness Index], and service and usage patterns are thought as the most important characteristics to solve this problem. QWE doubted that those customers with high CHI scores leave much, but those who are unhappy might leave, and so might those for whom CHI scores dropped recently. Also, number of support cases, average support priority, and usage information: logins, blogs, views, and days since last login are related with the customer retention.

3 Initial Data Analysis

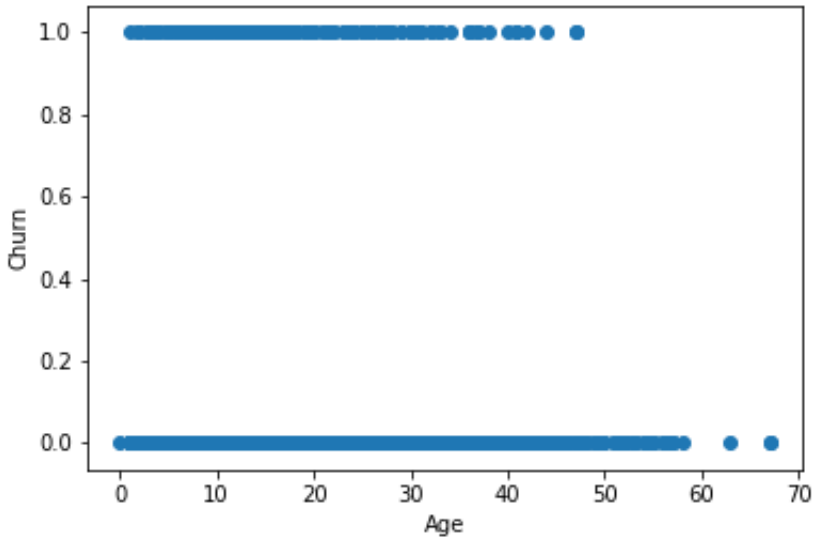
The dataset provided for this analysis includes 6,347 observations, each of which represents information for a given customer, across 13 variables:

- ID
- The id of a customer

- CustomerAgeinmonths
- ChurnYesNo
- CHIScoreMonth
- CHIScore
- SupportCasesMonth
- SupportCases
- SPMonth
- SP
- Logins
- BlogArticles
- Views
- DaysSinceLastLogin

4 Method And Results

4.1 Is Wall's belief about the dependence of churn rates on customer age supported by the data? To get some intuition, try visualizing this dependence (Hint: no need to run any statistical tests).



From the graph, we can not see any relationship between churn rate and customer age. As a result, customer age can not support the dependence of churn age.

4.2 To start, run a single regression model that best predicts the probability that a customer leaves.

To run a single regression model, we select the CHIScoreMonth to predict the customer churn because it has much higher correlation value than others, and also has a better p-value than others.

Characteristic	Correlation Value
ID	-0.106701
CustomerAgeinmonths	0.030215
ChurnYesNo	1.000000
CHIScoreMonth	-0.084005
CHIScore	-0.008713
SupportCasesMonth	-0.044973
SupportCases	-0.044407
SPMonth	-0.054935
SP	-0.019682
Logins	-0.043077
BlogArticles	-0.025090
Views	0.000007
DaysSinceLastLogin	0.111568

Characteristic	P-Value
CustomerAgeinmonths	1.60718369e-02
CHIScoreMonth	2.04157590e-11
CHIScore	4.87682982e-01
SupportCasesMonth	3.38343505e-04
SupportCases	4.01840246e-04
SPMonth	1.19213194e-05
SP	1.16910725e-01
Logins	5.97469659e-04
BlogArticles	4.56350308e-02
Views	9.99575334e-01
DaysSinceLastLogin	4.89975992e-19

We generate the single logistic regression model with argument class_weight='balanced' to handle the unbalanced data:

$$Equation : ChurnYesNo = -2.46064255 - 0.00615342 * CHIScoreMonth$$

$$Accuracy : 56.71971009925949\%$$

$$F - Score : 0.12208373282198785$$

a. What is the predicted probability that Customer 672 will leave between December 2011 and February 2012? Is that high or low? Did that customer actually leave? From the list below, we can see that the customer 672 actually did not leave.

ID	CHI Score	Probability	Leave
672	148	38.497424716712036%	NO

b. What about Customers 354 and 5,203? From the list below, we can see that the customer 354 and 5203 actually did not leave.

ID	CHI Score	Probability	Leave
354	139	39.8853787765764%	NO
5203	37	56.21420669369567%	NO

4.3 How sensible is the approach with a single model? Can you suggest a better approach?

For this single logistic model, when we want to know how sensible it is, we can calculate the F-Score of this model, and the F-Score is 0.12208373282198785.

If we want a better approach to predict the probability that a customer leaves, we can choose multiple logistic regreesion (MLR). In this case, we can use multiple customer characteristics to predict the probability that a customer leaves.

We choose the following five features in our model:

- CustomerAgeinmonths
- CHIScoreMonth0
- CHIScore01
- SP01
- DaysSinceLastLogin01

The reason why we choose this five features are based on the p-value calculate by multiple logistic model include all 11 feature (CustomerAgeinmonths, CHIScoreMonth0, CHIScore01, SupportCasesMonth0, SupportCases01, SPMonth0, SP01, Logins01, BlogArticles01, Views01, DaysSinceLastLogin01). The one's p-value is smaller, the more significant. (Our model data below)

Characteristic	P-Value
CustomerAgeinmonths	0.004
CHIScoreMonth	0.000
CHIScore	0.205
SupportCasesMonth	0.779
SupportCases	0.305
SPMonth	0.207
SP	0.163
Logins	0.959
BlogArticles	0.733
Views	0.322
DaysSinceLastLogin	0.000

After we select CustomerAgeinmonths, CHIScoreMonth0, CHIScore01, SP01, DaysSinceLastLogin01 as varibles in the model, we generate our Multiple Logistic Regression(MLR) with class_weight='balanced' :

Equation: $ChurnYesNo = -0.26734079 + 0.0120999 * CustomerAgeinmonths -$

$0.004228 * CHIScoreMonth0 + 0.00265245 * CHIScore01 + 0.04794592 * SP01 + 0.03206999 * DaysSinceLastLogin01$

$$Accuracy : 76.4140538837246\%$$

$$F - Score : 0.1859706362153344$$

Our MLR model is better than SLR model, with higher F-Score and Accuracy.

a. Provide updated estimates of probabilities that Customers 672, 354, and 5,203 will leave. Now, we can updated estimates of probabilities that Customers 672, 354, and 5203 will leave:

ID	Probability	Leave
672	34.68999537889548%	NO
354	35.69232880524198%	NO
5203	46.750860882514017%	NO

b. What factors contribute the most to the predicted probabilities that these customers will leave? Based on our multiple logistic model, the factors that contribute the most to predicted probabilities are following (with coefficient):

- SP01 (0.04794592)
- DaysSinceLastLogin01 (0.03206999)
- CustomerAgeinmonths (0.0120999)

We select this three factors because of their higher coefficient in our model.

4.4 Answer Wall's "ultimate question": provide the list of 100 customers with highest churn probabilities and the top three drivers of churn for each customer.

Here is the list of 100 customers with highest churn probabilities.

	ID	prob
1	2700	0.999999998393587
2	1496	0.9978747960857516
3	133	0.9977999106720837
4	2563	0.9975205803983748
5	1863	0.9973094157111098
6	1890	0.9931365584281898
7	1522	0.9928174161926085
8	871	0.992023749308499
9	52	0.9905019277291413
10	49	0.9868408109383346
11	2944	0.9804550240815016
12	1030	0.9801417910960859
13	1181	0.979543033460098
14	3257	0.9792703724089531
15	3686	0.9776174553714831

	ID	prob
16	3088	0.976094594377692
17	3027	0.9757418372662745
18	1108	0.9736529632177415
19	270	0.969917963226137
20	2281	0.9685991622461804
21	2079	0.9683580487887645
22	536	0.9675270695983309
23	1771	0.9663499831598592
24	194	0.9642226987502066
25	3581	0.9640512269580741
26	3583	0.9625830078724021
27	3293	0.9603936510302747
28	94	0.9601965173982313
29	192	0.9596037522990968
30	1219	0.9577904948543764
31	166	0.9518640044362658
32	1803	0.9450277822265417
33	2947	0.9429279400302903
34	2748	0.9427040716769648
35	4012	0.9410921097435994
36	2096	0.935777953608138
37	3787	0.9343552662128988
38	1010	0.932445343380797
39	1006	0.9313470559000587
40	110	0.9302158009050533
41	900	0.9291327628859228
42	927	0.9265184742663612
43	3468	0.9259860962987507
44	2033	0.9240091131060093
45	2011	0.9219280058137856
46	1948	0.9216844241089006
47	3695	0.9151235743135653
48	2740	0.9120427284787638
49	4025	0.9044357470830828
50	626	0.9043777016457477
51	1063	0.9018369354534261
52	3038	0.8967457420043813
53	2616	0.8943766643082819
54	863	0.8934986615508558
55	3647	0.8934964989654244
56	170	0.8925246806715886
57	1336	0.8904102095899663
58	935	0.8886157655290303
59	60	0.8879557173027112
60	3174	0.8841078394511775
61	1199	0.8834312687087589

	ID	prob
62	3467	0.8832392155449662
63	1802	0.8826784918442325
64	637	0.8764380069003616
65	111	0.8714893057639935
66	1316	0.8711734154841322
67	1672	0.8686213369811082
68	2992	0.8634193937042084
69	3333	0.8619611342178727
70	1921	0.8578886915794518
71	3759	0.8564897992842696
72	1198	0.8549243180260914
73	1106	0.8539523883958164
74	1501	0.8485481168595651
75	1127	0.847521491696026
76	2301	0.8465494676310138
77	381	0.844817937713592
78	1696	0.8389123400577667
79	2835	0.8353721284681913
80	631	0.8331971522730204
81	2651	0.8272950506236942
82	1	0.8231042413172577
83	901	0.8204406954984433
84	2611	0.8160908086992137
85	3541	0.8120995894022386
86	2028	0.811393113509938
87	1291	0.8106889392903045
88	3431	0.8066305131688226
89	14	0.8047927918720714
90	4194	0.8023247805508263
91	3	0.800962908977368
92	18	0.800962908977368
93	21	0.7990269005083666
94	4052	0.798899102269678
95	89	0.797866671861169
96	3254	0.7952179754541905
97	3803	0.7939412299540993
98	68	0.7929855674298147
99	51	0.7884313914887358
100	3522	0.7862275551348505

Based on the result of multiple logistic regression. We decide to determine the top three drivers of churn by the coefficient and the data set. According to their coefficient below:

Characteristic	Scores of features
CustomerAgeinmonths	0.0120999
CHIScoreMonth0	-0.004228

Characteristic	Scores of features
CHIScore01	0.00265245
SP01	0.04794592
DaysSinceLastLogin01	0.03206999

We select SP01, DaysSinceLastLogin01 and CustomerAgeinmonths as the top three drivers. However, the p-value of SP01 is much higher than CHIScoreMonth0, the fourth higher coefficient feature, which show that CHIScoreMonth0 is more significant than SP01, So we dicide to use CHIScoreMonth0 instead of SP01.

Finally, the top three drivers is:

	features
1	DaysSinceLastLogin01
2	CustomerAgeinmonths
3	CHIScoreMonth0

5 Conclusion

After in-depth analysis, the customer age seem like don't have relationship with the churn, but after determine our multiple logistic model with feature selection (Using p-Value), we find out that Customer Age have impact on churn with other features (showing in our MLR).

Our first model is single logistic regression model with CHIScoremonth0. We select our feature based on the Correlation of each features, and generate logistic regression model with it. Since the accuracy of our SLR model is only 56.7% and its F-Score is 0.12208, which is not sensible, we propose a multiple logistic regression model with five features(CustomerAgeinmonths, CHIScoreMonth0, CHIScore01, SP01, DaysSinceLastLogin01)

Our second model is multiple logistic regression model. We select our feature based on the p-Value calculated by MLR model with all 11 features, we decide to use CustomerAgeinmonths, CHIScoreMonth0, CHIScore01, SP01 and DaysSinceLast-Login01. And our MLR model's accuracy is 76.4% and its F-Score is 0.19, which is higher than our single logistic model with CHIScoremonth0.

According to our MLR model's coefficient and p-Value of each feature, we conclude that DaySinceLastLogin01, CustomerAgemonths and CHIScoreMonth0 are the top three drivers of customer leave. So QWE INC. should focus on this three factors.