PREDICTING CUSTOMER CHURN AT QWE INC.

Peilin Zhong, Zheng Li

March 9, 2019

1 Executive Summary

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

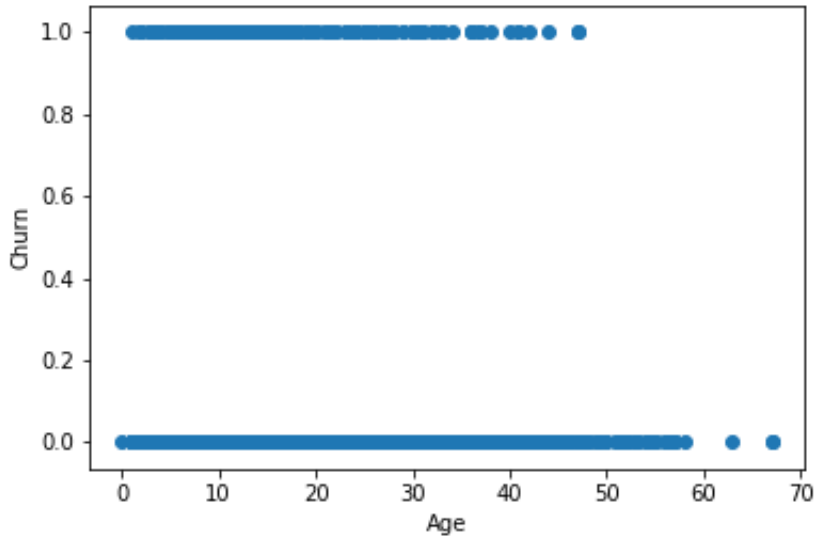
The dataset provided for this analysis includes 6,347 observations, each of which represents information for a given customer, across 13 variables:

- 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 Calue
ID	-0.106701
CustomerAgeinmonths	0.030215
ChurnYesNo	1.000000
CHIScoreMonth	-0.084005
CHIScore	-0.008713
SupportCasesMonth	-0.044973
SupportCases	-0.044407
SPMonth	-0.054935

Characteristic	Correlation Calue
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

Then, we can generate the single regression model:

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

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	3.3%	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	3.5%	NO
5203	37	6.36%	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.

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

a. Provide updated estimates of probabilities that Customers 672, 354, and 5,203 will leave. After we select CustomerAgeinmonths, CHIScoreMonth, CHIScore, SupportCasesMonth, SPMonth, DaysSinceLastLogin as varibles in the model, we can get the equation in the Multiple Logistic Regression(MLR):

$ChurnYesNo = -2.80257555 + 0.0155824 * CustomerAgeinmonths - 0.00609524 * CHIScoreMonth + 0.00339462 * CHIScore - 0.04366652 * SupportCasesMonth - 0.05840514 * SPMonth + 0.0109126 * DaysSinceLastLogin$

Now, we can updated estimates of probabilities that Customers 672, 354, and 5203 will leave:

ID	Probability	Leave
672	3.13%	NO
354	3.42%	NO
5203	5.48%	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 (0.02112769)
- CustomerAgeinmonths (0.01722109)
- DaysSinceLastLogin01 (0.0109209)

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	0.9706932532749759
2	1496	0.45860146666596513
3	133	0.4238869973237404
4	1863	0.3983660064529984
5	2563	0.3739013762334122
6	1890	0.3260457371423721
7	871	0.2944261628463586
8	1522	0.27630704086693697
9	49	0.2762175913513795
10	1181	0.2663714517863574
11	52	0.2657127481835317
12	1108	0.25918243789997775
13	94	0.25842687897598227
14	194	0.24627471305077495
15	3088	0.23899078031765428
16	2281	0.23330347160499162
17	110	0.21999437185327006
18	192	0.21337985722507477
19	1	0.20877373447395037
20	2944	0.20799899915779438
21	60	0.2023518135601883
22	166	0.20161062757817202
23	1030	0.19963298734468538
24	3257	0.19742445689456453
25	3581	0.1892020095223576
26	3027	0.18915248585955546
27	2011	0.1824382528893263
28	14	0.18174880003456853
29	3583	0.18099654907797935
30	270	0.1804255668760544
31	1803	0.1770662670778392
32	18	0.17668272094253745

	ID	prob
33	3	0.17668272094253745
34	536	0.17589175530764797
35	21	0.17419157029116547
36	2079	0.17135888936346835
37	1771	0.16766480371946627
38	2096	0.1668253487224701
39	1010	0.16523109521371426
40	1063	0.162979937973045
41	51	0.16121696047072587
42	3686	0.15938134827030853
43	89	0.15851758829231602
44	55	0.1556286224380185
45	59	0.15525368531612693
46	1219	0.1546081547305808
47	68	0.15402383693225655
48	61	0.15348590427170877
49	121	0.15300852395642026
50	3787	0.15295166753315856
51	1336	0.15293730553917975
52	95	0.15182680096947507
53	62	0.15173463635253776
54	137	0.15079003480689693
55	2748	0.1501817366718158
56	69	0.1495768821218133
57	154	0.14859806850094043
58	2	0.14811780953650047
59	1363	0.14805963301100536
60	109	0.14696358554126301
61	12	0.14649120168454716
62	171	0.14643247145057303
63	190	0.14643247145057303
64	119	0.14643247145057303
65	146	0.1456923881210843
66	3293	0.14523840243031713
67	42	0.1442931429822051
68	101	0.14429308601500176
69	2616	0.1437981484113255
70	183	0.14356464029732477
71	5	0.14206395232147404
72	2947	0.14076503650807626
73	1316	0.1404588834920363
74	863	0.14029576940629884
75	170	0.13962529942181123
76	1392	0.13878186168944595
77	1393	0.13851156103948142
78	1438	0.1380305656818362

	ID	prob
79	123	0.1376820393403658
80	156	0.1376213071362869
81	1006	0.1370481999529848
82	106	0.1367363462276371
83	1459	0.13637048967741422
84	2240	0.13634287040642043
85	76	0.13595341481233117
86	2992	0.1344066032790283
87	1672	0.13272880029654271
88	2835	0.13239219406530317
89	3333	0.13203315990186473
90	1395	0.1319979203288186
91	2235	0.1319979203288186
92	2255	0.1319979203288186
93	1478	0.1319979203288186
94	57	0.1318373540921706
95	16	0.13144035026577125
96	1143	0.13093931414947046
97	1405	0.1304134480838705
98	2244	0.13023715333288757
99	1378	0.13004484832262042
100	1488	0.13003730672203417

To find top three drivers of churn for each customer, we use f_regression to get the scores of features, and we can see that the top three customer characteristics are DaysSinceLastLogin, CHIScoreMonth and SPMonth.

Characteristic	Scores of features
CustomerAgeinmonths	5.79810180e+00
CHIScoreMonth	4.50934788e+01
CHIScore	4.81690527e-01
SupportCasesMonth	1.28592393e+01
SupportCases	1.25369732e+01
SPMonth	1.92063533e+01
SP	2.45889645e+00
Logins	1.17957239e+01
BlogArticles	3.99658854e+00
Views	2.83301624e-07
DaysSinceLastLogin	7.99745711e+01

5 Conclusion

After in-depth analysis of the