**Advanced Marketing Analytics (MKTG 881)**

**Assignment 3: Customer Churn at QWE**

**Jordan Martin**

**Due Date: 11/1 (Sunday) by 11:59 PM**

1. Summarize the issue that QWE faces.
   1. QWE is facing the issue of customers discontinuing service after their 6-month, 12-month, or month-to-month contract is concluded. When a customer calls to try to discontinue service, QWE, like most companies, offers discounts and special promotions to try to get the customer to keep using their services. This is obviously very expensive and if it could be avoided by making sure the customer doesn’t want to leave in the first place, it would decrease those costs and increase revenues as QWE would have less customers leaving. So, using their customer data, they’d like to create a model that can predict whether a given company will leave in the next one or two months so that they can proactively, and more cheaply, get them to stay with company rather than reactively and expensively do the same thing.
2. Is Wall’s belief about the dependence of ‘churn rates’ on ‘customer age with the company’ supported by the data? Provide evidence of the dependence through both data inspection (e.g., summary statistics, correlation, tabulation, but no regression) and visualization (e.g., a bar graph: churn vs. age). What are your observations? (Compare churn rate by age.)
   1. I think that for the most part, Wall is correct in his belief about churn rate being correlated with customer age with the company. From the two graphs below, you can see that both as a rate of churn and a count of churn, most companies do not end their contracts in the first five months, but several end theirs at 6 months and many end theirs at 12 months. The one point I would make to Wall is that there are still many customers that end their contracts all the way up to 18 months which takes away from his theory that 6 to 14 months are the most important. I think it’s also important to note that many companies end their contracts at 24 months as well, which would be two full 12 month contracts, but that rate decreases slightly at 36 and none ended theirs after 48, though there is no data of companies ending contracts at 48 so there may just not be enough historical data yet. Looking at correlation, in general, we see that there is very little correlation overall at about 3%, but for ages under 12 there is higher correlation 17%. I also tried doing a correlation for just ages divisible by six months and less than two years, but that also wasn’t significant. However, by generating a variable called sixmonth that was whether the age is divisible by six, I found 10% correlation to churn. Similarly, I found a correlation of 14% for ages that are divisible by 12. This means that Wall is probably onto something when he thought that customers are commonly cancelling at the end of their contracts. Finally, when looking at the correlation of total churns by age, I found that there was a relatively high negative correlation at -19%, but under 13 months, it’s highly positively correlated at 65%.
   2. Churn Rate by Age:
      1. 
   3. Churn Count by Age
      1. 
   4. Correlation of age and churn
      1. 
   5. Correlation of age and churn for age less than or equal to 12 months
      1. 
   6. Correlation of age and churn for age divisible by 6 months and less than 2 years
      1. 
   7. Correlation of age divisible by 6 and churn
      1. 
   8. Correlation of age divisible by 12 and churn
      1. 
   9. Correlation of age to total churns
      1. 
   10. Correlation of age to total churns for age less than 13 months
       1. 
3. 1. *Specify* and run a logit model (**Aggregate Model**) that predicts the probability that a customer leaves (Note: use βs when specifying your model). Show the estimation results by plugging the coefficients into your model and *interpret* the meaning of each coefficient.
      1. Here are the results of the logit model:
         1. 
      2. After removing the statistically insignificant variables, our results become as follows:
         1. 
      3. Because this is a logistic regression, the coefficients are multiplied by their values, then the constant is added. This value, we can call x, is then plugged into the following formula: Probability = 1/(1+exp(-x))
      4. So if we look at an example, say an account that is 9 months old, had a customer happiness index of 250 this month, which was 150 less than last month, had viewed 20 more blog articles than last month and last logged in 10 days ago:
         1. We’d get x = 0.0149233\*9 + (-0.0055843)\*250 + (0.0094253)\*150 + (-0.0001133)\*20 + 0.017115\*10 + (-2.784868) = -2.4639543
         2. And exp(x) = exp(-2.4639543) = 0.08509, so there is an 8.5 percent chance of this particular customer churning.
      5. In general, a higher positive value means that a higher value of that variable increases the chance of churn and a higher negative value of that variable decreases the chance of churn.
   2. What is the overall correct classification rate of the model? (change the cutoff point if necessary such as *estat classification, cutoff(0.2)*)
      1. I tried using a few different cutoffs, but based on the raw data, there was not a very good model that I could construct. The best was probably a cutoff of 0.2 which yielded the following result:
         1. 
      2. This is just 94.83% correct, but for actual churns, it was only 14.29% correct, so that overall classification rate is very misleading.
4. Why does the churn rate spike at 12 months? How would you control for this spike in your model?
   1. The churn rate spikes at 12 months because a lot of contracts are year-long, so after the first year a lot of companies decide that they no longer want or need QWE’s services, so they let the contract expire without renewing.
   2. To control for this spike, I would add a dummy variable that is 1 if the age is 12 and 0 otherwise. It may also be worth looking at the smaller spikes at 6, 18, and 24 as I mentioned earlier.
5. 1. To try an alternative approach, divide the dataset into three groups using AGE (i.e., how long a customer has been with QWE). Explain how you divide customers into three age groups and why. (To solve this question, you may want to split the original data into three datasets or you can use the ‘if-statements’ when estimating each data)
      1. To divide the dataset into three groups by age, I ran the following commands:
         1. gen agegr=1 if age<=6
            1. This sets a dummy variable agegr to 1 if the age is less than or equal to six months
         2. replace agegr=2 if age>6 & age<=12
            1. This sets the dummy variable agegr to 2 if the age is between 7 and 12 inclusively
         3. replace agegr=3 if age>12
            1. This sets the dummy variable agegr to 3 if the age is greater than 12
      2. This is useful because as we noticed before, the different values for age affect the rate of churn differently. In the first six months, very few companies churn, but between 6 and 12, and particularly at 6 and 12, there is a lot of churn and then churn decreases as we pass the one year mark.
   2. Estimate three logit models based on the AGE groups (**Separate Models**) and show the results. (Note: create the dummy variable of AGE=12 if the dataset includes AGE=12 months and include this dummy variable in the model.)
      1. For age less than 6 months:
         1. 
      2. For age 6 to 12 months:
         1. 
      3. For age greater than 12 months:
         1. 
   3. Explain your findings about the relation between churn and age in each model (**Separate Models**) compared to the relation in the **Aggregate Model** in Q3 (e.g., significance, sign, magnitude).
      1. For the first age group, the churn rate is extremely low below 3%, so the model didn’t predict any to churn even with a cutoff of 0.2. For this one, the age has the same significance as the aggregate model, with a p-value of 0.004. The sign is the same, but the magnitude is about three times the size, so age plays more of a factor of increasing churn in the first six years than overall for the full dataset.
      2. For the second age group, the churn rate is the highest of any age group at 7.3% and the model was able to correctly classify 35% of churns which is much better than the aggregate model. For this one, the age is hardly significant at all with a p-value of 0.174, but the dummy variable age12 which is for accounts that are exactly 12 months old, is extremely significant with a p-value of 0.000. Age, overall plays a negative role in prediction this time as the sign is flipped, but the magnitude is about the same as the aggregate model. That being said, it doesn’t really matter in the model; what matters is the age12 dummy variable. This has a large positive value, 2.16, so being 12 months old correlates to a large increase in the probability of churn which is exactly what we expected.
      3. For the last age group, the churn rate is also fairly low and the model didn’t predict any of the companies to have a greater than 50% chance of churning. Age is very significant in this model with a p-value of 0.000, but the sign is still negative so as age goes up, the likelihood of churn goes down. It’s less of a change than the aggregate model though.
6. Answer Wall’s ultimate question: Who are the top 100 customers with the highest churn probabilities? You need to use the estimation results of the Separate Models to calculate the probabilities.
   1. List the top 10 customers with their IDs and their age group.
      1. 357
      2. 335
      3. 279
      4. 4500
      5. 317
      6. 257
      7. 371
      8. 3313
      9. 488
      10. 543
   2. Compare the actual churn rate of these 100 customers with the overall churn rate in the whole data and explain why the two rates are different.
      1. The overall churn rate in the whole data is about 5% while the churn rate for the top 10 customers with highest churn probability was 38%. This is much better than the aggregate model which only had a churn probability of 13%.
      2. These two rates are different because our model does a better job of predicting which customers will churn than a model that just randomly picks them. In the top 100, the probability ranges from 64% at the high end down to 23% at the low end. While the model can’t conclusively say that any of these customers will churn, it does tell us that they have a good chance to churn, so QWE would be wise to try to proactively work to try to keep these customers happy as they are the likeliest to leave in the next two months.