

Predicting New Customer Retention for Online Dieting & Fitness Programs



December 11, 2007

BUDT733 DC01

Team Four

Amy Brunner

Harin Sandhoo

Lilah Pomerance

Paola Nasser

Srinath Bala

Executive Summary

GymAmerica.com is an online personal fitness and diet website where customers can obtain, as well as customize, a personalized workout routine and diet plan. The goal of this project is to predict whether a new user will become a paying customer for Gymamerica.com, after using the 10-day free trial period. The data we gathered were for customers who have registered between January 1, 2004 and June 30, 2007, have entered their credit card, have selected a subscription to either exercise and/or diet service, and who have entered the free trial after finishing configuration of the exercise and/or diet application.

The original data set consisted of over 24,379 observations and 37 variables (categorical and numerical). After several exploratory studies and with the help of domain knowledge, the final predictor list was narrowed to the following variables: gender, squat preference, strength ability, age, strength workout location, strength workout length, workout plan, and diet plan. Modeling with the final predictors concluded that the customers who are more likely to pay after the free trial are men with only an exercise program who like to do barbell squats, like to work out at home or at the gym, and like to use both free weights and machinery. In addition it was also found that customers that selected the “Advanced Toning for Women Workout” are more likely to pay than any other workout plan.

Therefore, it is recommended that GymAmerica.com pursue a marketing campaign targeted towards individuals who fit the above profile. This model could be used to send targeted email and mail offers to the customers who are more likely to pay after they fill in a customer profile. Another recommendation is to change the workout plan options webpage, so that if the customer is a female, the default workout selected is “Advanced Toning for Women Workout”, which will increase the odds that this user will pay. It is also suggested that the company improve how customer configuration preferences are stored in the database, and how questions are asked to improve the quality of future statistical models. Specifically GymAmerica.com should change how age is asked (from date of birth to a number) and ensure that the history of all preferences chosen is stored, to avoid overriding first time preferences (like strength and cardio experience).

Technical Summary

Data collection and cleanup was not trivial. Though one of the project team members had access to all of Gymamerica's SQL Server databases, knowing what data to collect and how to collect it was a challenge. The starting point was the collection of all variables that a customer enters during exercise and diet registration, as well as any billing details and demographics entered in the main registration. Initial results showed over 24,000 records with many missing values. After examining the data, a discrepancy was identified in the age field where it appeared that too many records had an age of 18 (the default age). The issue was traced to data entries from customers who had not finished entering configuration preferences. To reduce the scope of the analysis, and concentrate only on the customers who have seen the application, such records were removed. Also filtered out were any test customers, any customers who have selected only the diet application (because they were less than 1% and had missing values for all of the exercise preferences), any international customers except for Canada, and any customer who have been transferred from a previous Genesant corporate partner.

Another problem was that some customers' preferences had been archived and required restoration and access to the archived data. In addition, some workout preferences such as workout length and number of workouts per week were missing for customers who did not select the "Select Workout Plan". This data had to be inferred from the workout plan description. At the end of this process, there were 10,431 records, and 37 variables. During data preprocessing, new variables were derived from existing ones, while removing the redundant ones. For example, from zip code, state, time zone and country, the location region was derived. Height was removed because it was redundant (BMI, Body mass index is a function of weight and height).

Data Analysis included several exploratory tests including Chi-Square test, Statistical Average Analysis (comparing averages of the output variable for each predictor), Correlation, Bar Charts and Box Plots, with a strong influence of domain knowledge guiding the effort. Some of the charts explored are shown in Exhibit A. Based on their significance across in these tests, a set of 15 predictors were chosen.¹ Strength Experience was one of the predictors that was most

¹ *StrengthWorkoutLocation; Log_StrengthAbility; Gender; DietConveniencePlan; DietPlan; WeightGoal; BMI_bins; StrengthWorkoutLength; SquatPreference; Goal; Age_Bins_2; WorkoutPlan; HasInjury; Weight_Bins; HasSpecialOffer*

significant in the tests, but there were problems found in how this data had been stored in the database. Because of the questionable reliability of the data, we had to exclude it from the list.

Since the predictors were primarily categorical, they were converted to dummy variables, which led to 54 predictors. To reduce these predictors, the bins of the numerical variables (age, weight and strength ability) were un-binned, leaving them as numerical (except for the Naïve Bayes model). To further reduce the number of dummy variables, pivot tables were used. These pivot tables listed the original variables with the average of the Y-variable for each of the categories (see Exhibit B). The categories that appeared similar to each other with respect to the distribution of the output variable were combined.

Five different models were run, due to the predictive nature of this project: Classification Tree, Discriminant Analysis, Logistic Regression, K Nearest Neighbors, and Naïve Bayes. Comparing the results to the Naïve rule (error rate of 40.95%), the models showed high error rates (38% - 46%), and thus warranted further examination. Some non-significant variables were removed, and some more dummy variables were combined after examining the p-values and the DA scores (see Exhibit C for comparison of 5 models). The chosen model, based on the best validation and test error rate, as well as the sensitivity and specificity of the test set, was the logistic regression (See Exhibit D for output). The significant predictors included: male, squat preference, no diet plan selected, strength ability, age, workout location, workout length, workout plan, and diet plan.

We recommend Gymamerica to pursue a marketing campaign targeted towards individuals who are male, who like to do barbell squats, and like to workout with both free weights and machinery. We also recommend them to gather information about the people who use the fitness profile, but fail to enter the free trial, and use it as an input to the logistic regression model. For the customers that the model predicts are going to become paying customers, Gymamerica should send a special offer or an extra incentive for them to come back to the site and enter the free trial. In addition we recommend them to make the “Advanced Toning for Women” the default option for women. Also, based on the efforts during the data collection activities, it is highly recommended that Gymamerica store user configurations more accurately in the database. This will ensure that any useful data collected by the company can be greatly beneficial for any future modeling.

Exhibit A: Data Exploration Graphs

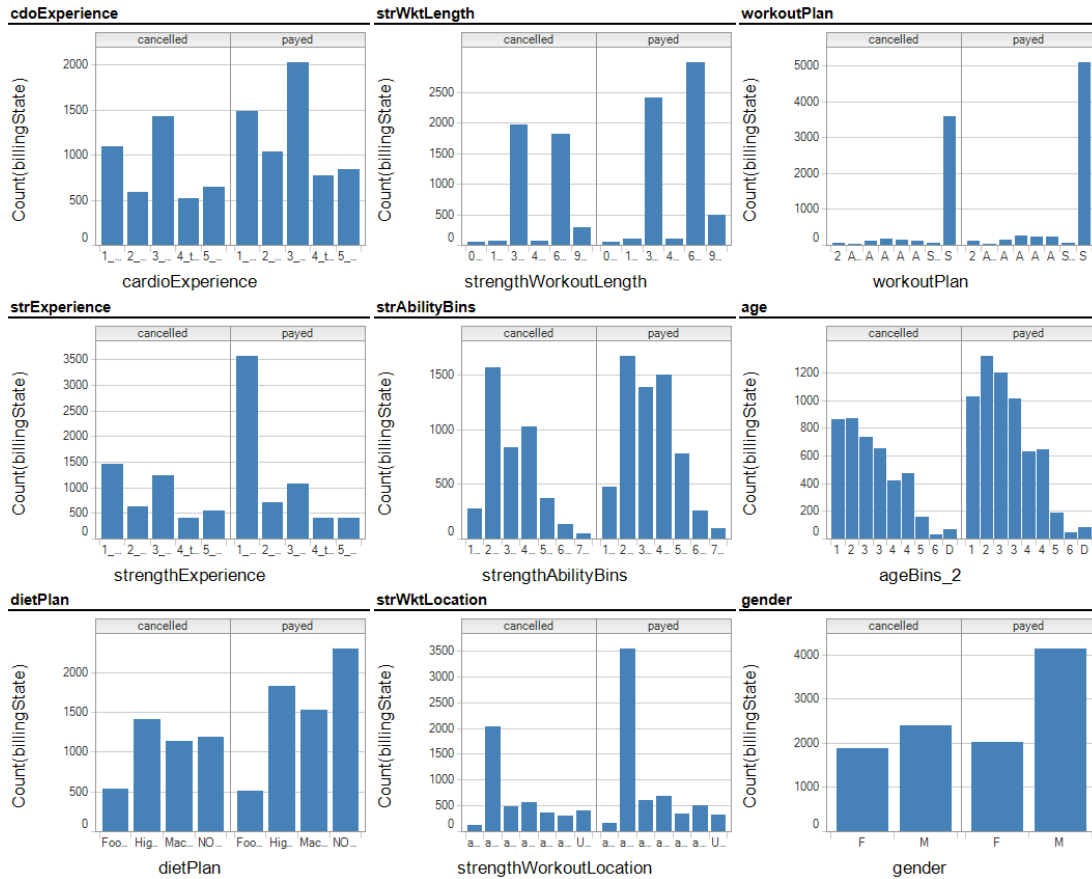


Exhibit B: Predictor Analysis Pivot Tables

| workoutPlan | count | avg |
|---------------------------------|-------|----------|
| Advanced 5K w/ Weight Training | 33 | 0.636364 |
| Advanced Toning for Men | 381 | 0.635171 |
| Advanced Power Workout | 427 | 0.59719 |
| Select Cardio Only Workout Plan | 113 | 0.504425 |
| 21-Day Advanced Core Toning | 169 | 0.615385 |
| Advanced Circuit Training | 275 | 0.552727 |
| Advanced Toning for Women | 340 | 0.697059 |
| Select Workout Plan | 8693 | 0.585759 |

| dietPlan | count | avg |
|--------------------|-------|----------|
| Food Pyramid Diet | 1044 | 0.484674 |
| High Protein Diet | 3239 | 0.563754 |
| NO DIET | 3481 | 0.659581 |
| Macronutrient Diet | 2667 | 0.574428 |

| weightGoal | count | avg |
|-----------------|-------|----------|
| Lose Weight | 5800 | 0.547759 |
| Maintain Weight | 294 | 0.540816 |
| NO DIET | 3481 | 0.659581 |
| Gain Weight | 856 | 0.616822 |

| strWktLength | count | avg |
|--------------|-------|----------|
| 30 minutes | 4383 | 0.549624 |
| 45 minutes | 169 | 0.615385 |
| 0 minutes | 113 | 0.504425 |
| 120 minutes | 166 | 0.614458 |
| 60 minutes | 4814 | 0.620897 |
| 90 minutes | 786 | 0.63486 |

| dietConveniencePlan | count | avg |
|----------------------------|-------|----------|
| The On-The-Go Meal Plan | 2312 | 0.597751 |
| The Combination Meal Plan | 2532 | 0.553318 |
| The Home Cooking Meal Plan | 443 | 0.555305 |
| The Easy Home Meal Plan | 1659 | 0.500904 |
| NO DIET | 3485 | 0.659971 |

Exhibit C: Gymamerica.com Modeling Results

| | Training Set Error% | Validation Set Error% | Test Set Error% | Test Set Sensitivity | Test Set Specificity |
|-----------------------|---------------------|-----------------------|-----------------|----------------------|----------------------|
| Logistic Regression | 37.33% | 39.76% | 37.20% | 86.24% | 28.23% |
| KNN | 35.01% | 41.48% | 37.44% | 85.68% | 28.47% |
| Classification Tree | 38.90% | 41.58% | 38.21% | 89.38% | 18.62% |
| Naïve Bayes | 38.88% | 40.91% | 38.83% | 78.12% | 36.18% |
| Discriminant Analysis | 40.43% | 42.22% | 40.17% | 63.64% | 54.21% |
| Naïve Rule | 40.95% | | | | |

Exhibit D: Logistic Regression Model

Prior class probabilities

| According to relative occurrences in training data | |
|--|-------------------------------|
| Class | Prob. |
| payed | 0.598542945 <-- Success Class |
| cancelled | 0.401457055 |

The Regression Model

| Input variables | Coefficient | Std. Error | p-value | Odds |
|---------------------------------------|-------------|------------|------------|------------|
| Constant term | -0.24233337 | 0.48623437 | 0.61821061 | - |
| gender_M | 0.63393086 | 0.08208836 | 0 | 1.88500571 |
| squatPreference_yes | 0.23397914 | 0.06712396 | 0.00049071 | 1.26361811 |
| log(strAbility) | -0.23567867 | 0.06456982 | 0.00026226 | 0.79003447 |
| log(age) | 0.14171952 | 0.10306153 | 0.16910163 | 1.15225339 |
| strWktLoc_atGymFreeWts | 0.70793337 | 0.22595046 | 0.0017295 | 2.02979207 |
| strWktLoc_atGymFreeWts&Machines | 1.35648978 | 0.15434419 | 0 | 3.8825407 |
| strWktLoc_atGymMachines | 1.11714578 | 0.17428556 | 0 | 3.05611873 |
| strWktLoc_atHomeDumbbells | 1.22130656 | 0.17483339 | 0 | 3.39161611 |
| strWktLoc_atHomeFreeWts | 0.80089557 | 0.18768552 | 0.00001979 | 2.22753501 |
| strWktLoc_atHomeFreeWts&LegExtMachine | 1.31087041 | 0.18496068 | 0 | 3.70940113 |
| strWktLen_45 minutes or more | 0.10658011 | 0.06212238 | 0.08622657 | 1.11246705 |
| wktPlan_Advanced Toning for Women | 0.70400673 | 0.20847158 | 0.00073283 | 2.02183747 |
| wktPlan_Select Workout Plan | -0.61317694 | 0.11918895 | 0.00000027 | 0.54162741 |
| dietPlan_NO DIET | 0.31598803 | 0.06445931 | 0.00000095 | 1.37161386 |
| dietPlan_PyramidDiet | -0.11608823 | 0.09830552 | 0.23764549 | 0.89039665 |

| | |
|----------------------------|-------------|
| Residual df | 5200 |
| Residual Dev. | 6747.594727 |
| % Success in training data | 59.85429448 |
| # Iterations used | 10 |
| Multiple R-squared | 0.03975823 |

Training Data scoring - Summary Report

| Cut off Prob.Val. for Success (Updatable) | | 0.5 | |
|---|-----------------|-----------|---------|
| Classification Confusion Matrix | | | |
| | Predicted Class | | |
| Actual Class | payed | cancelled | |
| payed | 2675 | 447 | |
| cancelled | 1500 | 594 | |
| Error Report | | | |
| Class | # Cases | # Errors | % Error |
| payed | 3122 | 447 | 14.32 |
| cancelled | 2094 | 1500 | 71.63 |
| Overall | 5216 | 1947 | 37.33 |

Validation Data scoring - Summary Report

| Cut off Prob.Val. for Success (Updatable) | | 0.5 | |
|---|-----------------|-----------|---------|
| Classification Confusion Matrix | | | |
| | Predicted Class | | |
| Actual Class | payed | cancelled | |
| payed | 1516 | 279 | |
| cancelled | 965 | 369 | |
| Error Report | | | |
| Class | # Cases | # Errors | % Error |
| payed | 1795 | 279 | 15.54 |
| cancelled | 1334 | 965 | 72.34 |
| Overall | 3129 | 1244 | 39.76 |

Test Data scoring - Summary Report

| Cut off Prob.Val. for Success (Updatable) | | 0.5 | |
|---|-----------------|-----------|---------|
| Classification Confusion Matrix | | | |
| | Predicted Class | | |
| Actual Class | payed | cancelled | |
| payed | 1072 | 171 | |
| cancelled | 605 | 238 | |
| Error Report | | | |
| Class | # Cases | # Errors | % Error |
| payed | 1243 | 171 | 13.76 |
| cancelled | 843 | 605 | 71.77 |
| Overall | 2086 | 776 | 37.20 |

| | |
|-------------|--------|
| Sensitivity | 86.24% |
| Specificity | 28.23% |