Final Project: Data Analysis for Decision Makers (BUDT733), Fall 2007

Population Analysis for Member Retention:

Assisting the Association of Financial Professionals

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

Signature	Typed Name
	Cytena Hubbard
	Christopher Klejdys
	Galina Kozachenko
	Joshua Miller
	Ahsanur Rahman

Executive Summary

Our goal is to classify (i.e. predict) renewals for the Association of Financial Professionals (AFP) in order to decrease marketing costs and increase return on investment for membership retention campaigns. AFP staff has profiled membership patterns and identified first-year members as a "highly probable to not-renew group." AFP came up with some basic profiles for the new members, and the "caution" factors that separate non-renewals from renewals. On average, higher activity (participation in the newsletter and discussion list) resulted in higher renewal rates. We calculate the costs to the firm of using the predictive results from the different models but we do not make a specific recommendation of what model to use based on the various costs because there are tradeoffs in terms of financial costs and model accuracy.

AFP revenue is $760 per member (not including sponsorship) on average: $395 of that were member dues and the rest was the product of average annual events costs and member participation rates. AFP has estimated that new campaign "member retention" would cost them $36/member due to the cost of the telemarketing call, one email and one direct mailing). AFP understands that a certain percentage of non-renewals are natural to any Association.

AFP provided our team with data that included a membership snapshot for 2006 along with most of the variables from 2005 profile research which includes two additional variables not provided in the 2006 data that measure the certification cycle for members. AFP was not able to provide 2 variables that we thought would be important to predict renewals: last date the member accessed the web and last date a member made a purchase online. See Attachments 4 for the initial list of variables provide by AFP and attachment 5 for the description of how some of the original variables were transformed to better suit our data processing needs.

We built several models for the associates and practitioners data sets using K-Nearest Neighbors, Logistic Regression, Naïve Bayes, and Discriminant Analysis. When studying the data, we identified several variables that clearly had different metrics for renewals and non-renewals. The classification tree indicated that the longer the member stays with AFP, the higher the chances are that it is a renewing member (See Attachment 1, "classification function").

The technical summary includes the metrics for the 4 best models for practitioners and associates. We provide the results of the sensitivity, specificity, false positive, false negative and overall error for the models in the attachments for this report. To understand the full effect of the models, we also computed business measures: potential profit, upfront cost and ROI.

Logistic regression and discriminant analysis have the lowest overall error for the models we choose, but they come second when detecting the important class (non-renewals) as compared to Naïve Bayes (See Exhibits 2 and 3). Use of the logistic regression and discriminant analysis models for prediction would also yield the highest return on investment because of the lower upfront cost of the predictions as these models classify less number of members as non-renews, and less money would be spent on the "member retention" marketing campaign in this case. If AFP's goal is minimize the variable costs of a membership campaign, they should go with one of these two models.

Logistic regression may be a better model of the two, since it can also serve as an explanatory model. Logistic regression may be used in the future to "test" the model by adding new variables to see if the predictive ability will improve. However, if AFP's goal is to increase membership by increasing the renewal rates regardless of variable costs, AFP's best choice is Naïve Bayes, since it has the highest predictive ability in regards to non-renewing members. We believe that the Naïve Bayes model for Practitioners has a higher predictive ability for non-renewals because there are less "accidental" practitioners than associates (i.e. members misclassifying themselves), and there is a higher misclassification cost attached to the practitioners than associates.

Technical Summary

AFP data has 15042 observations that represent a snapshot of AFP members in 2006. 20.4 percent of the members did not renew their membership into 2007 (nonrenewals) and this is our "success" class that needs to be predicted. The raw data sample can be viewed in the attachment 1. The data comes from the AFP database DMG, and there is no problem with future availability of the data. AFP did not provide us, however, with two variables that could have been useful: LAST ONLINE ACTIVITY DATE and LAST ONLINE ORDER DATE.

We began our analysis by checking the data for quality. The raw data has 14 variables. Most of the variables are categorical, and only "YEARS WITH AFP" and "YEARS SINCE DELINQEUNCY" are numerical. There was a high percentage of missing data in the variable "INDUSTRY", so we created a new industry code to capture this called "blank". There are many observations that have no data for the TITLE CODE, so we replaced the blank data with a 0.

Some of the variables are categorical with several classes and needed to be binned. We binned the INDUSTRY variable with 24 classes into 10 bins; the binning is done based on domain knowledge, renewal rate and size of industry in relations to the population: for example, we combined Petroleum, Energy and Utility industries into a bin called Energy – these industries have a similar renewal rate, population size and they are somewhat close to each other by nature. We binned the Variable TITLE_CODE into 6 bins – codes 1, 2 and 3 (Treasurer, Assistant Treasurer and CFO) we put into a bin 1&2&3, and members with no codes into a bin NO CODE, based on the renewal patterns for these title codes (based on domain knowledge).

We created 2 dummy variables to replace MEMBER TYPE GROUP:

1. TYPE where 1 was corporate practitioners and 0 was associates, and
2. COUNTRY where 1 was USA and 0 was Canada.

Variable TITLE GROUP was converted into a dummy variable TITLE with 1 representing "senior" titles (Treasurer, CFO, etc.) and 0 representing "core" titles (Manager, Accountant).

Further study of the data revealed that non-renewals had a much lower participation in discussion lists and newsletters. The other variables that separated non-renewing from renewing members were: Certification variables (CTP cycle end YES and CTP cycle end NO), NO CODE (there were many more members without a title code that did not renew), TYPE (there were 10% less practitioners who did not renew). After we ran a classification tree on the data, we confirmed that some of these were useful for classification. The regression tree put the variables in the following hierarchical order: YEARS WITH AFP, NEWSLETTER, CTP-CYCLE_END_NO and NO CODE.

The fact that there seemed to be fewer practitioners who were non-renewals caused us to look at non-renewals for practitioners and associates (vendors) separately (Attachment 1, 2.a). Indeed, the renewal rates were different – 19% for practitioners, and 23% for associates. Non-renewing practitioners had 6% less members with no title code, 11% less members with senior codes, 10% less members who participated in the Annual conference (that can be explained by "accidental" associates members who work the booth on the exhibit floor), and 7% less members who hold AFP certification. 50% of non-renewing practitioners have been with AFP for at least 2 years, while 50% of associates have been with AFP for only 1 year or less.

After talking with the Director of Research and Data Standards about the misclassification cost, we confirmed that it was more important to correctly detect non-renewing practitioners than associates. The misclassification ratios AFP gave us were: 5 to 1 for practitioners and 4 to 1 for associates. The different misclassification costs led us to alter the cutoff probability values for the models. Based on our analysis, the optimal cutoff value for the models hovered around .2 with success defined as “nonrenewal.”

Based on what we learned from the Director of Research and Data Standards, we separated the data into two data sets – Associates and Practitioners. Our goal was to predict the non-renewals; our large data sets gave us greater flexibility in choice of predictive methods. We decided to try the following methods: Classification tree, KNN, Naïve Bayes, Discriminant Analysis and Logistic Regression. We made some adjustments to the data in order to use many of the above models: additional binning of the numeric variables to be able to run Naïve Bayes, and normalizing the numerical data to run Discriminant Analysis. We also ran a correlation matrix, and saw that our variable TITLE is highly correlated with the binned variables for Title codes.

We selected what we believed to be the most predictive variables and introduced them into the Logistic regression on the data set that included both associates and practitioners. Interestingly, we did not end up using the “YEARS FROM DELINQUENCY” variable in our final analysis because the logistic regression model returned a high p-value for it. However, the Classification Tree method indicated that this variable was more important than the logistic regression method.

We re-computed the “YEARS FROM DELINQUENCY” variable into a categorical one: 0 if a member was not delinquent, 1 if a member was delinquent. After we did this, the variable had a high p-value. After running several logistic regression models, we identified the number of variables that were good predictors. We ran several Logistic regressions on the Associates and Practitioners data sets, and altered the variables included in the model for two these two data sets based on the results of the regressions. The Associates model had a high p-value for BLANKUNEMPLOYED, while the Practitioners models had a low p-value for that same variable. This indicates that while it was important for practitioners whether a member did not indicate an industry or was unemployed, it did not matter much for the associates non-renewals prediction.

When we ran all of the models, we received the best overall predictive results from the Discriminant Analysis in terms of overall error and return on investment (see below). The reason we received better results was because we were able to input the misclassification cost for this model which resulted in a much better prediction of the “success” class – non-renewals. Other models had a very high error when it came to predicting non-renewals, and very low error when it came to predicting renewals (See Attachments 2 and 3). We had to introduce the misclassification cost into the other models – we computed the expected cost and best cut off value for each of the models, taking into consideration the cost of renewals, non-renewals, probability of non-renewing and sensitivity/specificity errors.

At the end, we obtained 4 predictive models for each data set.

Corporate Practitioners

	Profit	Upfront cost	ROI	Sensitivity	Specificity	False Pos	False Neg	Overall	Exp. Cost
LR	$26,448.00	$28,728.00	92%	71%	63%	70%	10%	35.46%	1.21
DA	$26,448.00	$28,044.00	94%	70%	64%	69%	10%	34.76%	1.21
KNN	$25,872.00	$29,304.00	88%	71%	62%	70%	10%	36.32%	1.20
NNB	$26,292.00	$34,812.00	76%	78%	54%	72%	8%	41.78%	1.21

Associates

	profit	upfront cost	ROI	Sensitivity	Specificity	False Positive Error	False neg	Overall	Exp Cost
LR	$20,640.00	$16,524.00	125%	61%	67%	64%	15%	34.60%	1.06

DA	$20,808.00	$16,128.00	129%	60%	68%	64%	15%	33.82%	1.06
KNN	$22,032.00	$21,060.00	105%	71%	56%	68%	14%	40.98%	1.05
NNB	$23,616.00	$21,528.00	110%	74%	55%	67%	12%	40.55%	1.08

Recommendations:

1) Go with Logistic Regression if the goal is to minimize the upfront cost and maximize the ROI for the Associates
2) Go with Discriminant Analysis if the goal is to minimize upfront cost and ROI for practitioners
3) Go with the Naïve Bayse if the goal is to increase the number of renewals
4) When the other two variables become available after the re-design of the web site and web store, introduce two new variables into the models to observe whether the predictive ability of the models increase.