Model-based Sampling Using an Online Probability Panel

Meimeizi Zhu, Angela Fontes, Justine Bulgar-Medina
NORC at the University of Chicago, 55 E Monroe St, 30th Floor, Chicago, IL 60603

Abstract
In order to help researchers better understand individual financial wellbeing and management, a pilot study was launched in September 2019. The study invited selected AmeriSpeak panelists to enroll in an online application to link their financial institutions or accounts. This allows researchers to passively track all account transactions in real time. In this present work, we conducted a model-based sampling method to select samples for the full launch of financial transaction study in 2020. A model has been built to predict response rate, consent rate, and enrollment rate for the rest of the panelists by analyzing the pilot data from hypothesized to have primary influence include the panelist's demographic information, incentive levels, and financial profile information. Ultimately, panelists who have a higher predicted response rate, consent rate or enrollment rate were oversampled and invited to the pilot study. The predictors participate with the intention of maximizing BAA enrollments.

Key Words: probability sample, model-based sampling, online panel

1. Introduction

In September 2019, we launched a pilot study which invited selected AmeriSpeak panelists to enroll in an online application to link their financial institutions or accounts. This study helps researchers better understand individual financial wellbeing and management, and also allows researchers to passively track all account transactions in real time.

In this paper, we conducted a model-based sampling method to select samples for the full launch of financial transaction study in 2020. A model has been built to predict response rate, consent rate, and enrollment rate for the rest of the panelists by analyzing the pilot data from hypothesized to have primary influence include the panelist's demographic information, incentive levels, and financial profile information. Ultimately, panelists who have a higher predicted response rate, consent rate or enrollment rate were oversampled and invited to the pilot study. The predictors participate with the intention of maximizing BAA enrollments.

Data for this study are collected using NORC's AmeriSpeak Panel. AmeriSpeak® is the first U.S. multi-client household panel to combine the speed and cost-effectiveness of panel surveys with enhanced representativeness of the U.S. population, an approach designed to achieve an industry-leading response rate. Developed and funded by NORC, AmeriSpeak is the most scientifically rigorous panel solution available in the U.S., and gives NORC clients a breakthrough option for conducting statistical surveys of the population. Spending data (FINData) are collected on select AmeriSpeak panelists and track transaction level data on all accounts linked by the panelist. Transactions are categorized in to one of 13 expenditure categories.

2. Models for Imbalanced Data

Imbalanced data refers to a classification problem where the number of observations per class is not equally distributed. There are multiple ways to deal with imbalanced data but we focused on the following models in this paper:

- Logistic regression: This is the baseline model for comparison.
- Synthetic Minority Over-sampling Technique (SMOTE): This is a technique that generates new observations by interpolating between observations in the original dataset.
- K-Nearest Neighbors (KNN): This is a nonparametric method, which does not require any prior knowledge of the distribution.
- Random forest: R package "randomForest" was applied and option "*sampsize*" was used to balance data.
- Xgboost: R package "xgboost" was applied and option "*scale_pos_weight*" (ratio of number of negative class to the positive class) was used to balance data.

3. Model Comparisons

3.1 Identify Potential Enrollments

In the pilot study, we sent out 921 invitations and 872 of them completed financial well-being survey. Among the 872 respondents, less than 10% of them actually enrolled in our online application and linked their financial institutions or accounts. See Figure 1.

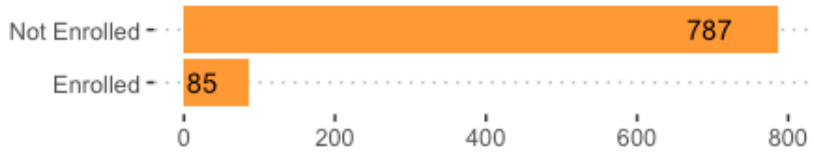

Figure 1: Plot for distribution of enrollment

We then produced a cluster plot for the enrollments and non-enrollments. As you can see from Figure 2 below, the two groups, show little differentiation on the principal components, which suggests that classifying data into the groups and predicting enrollments could be difficult.

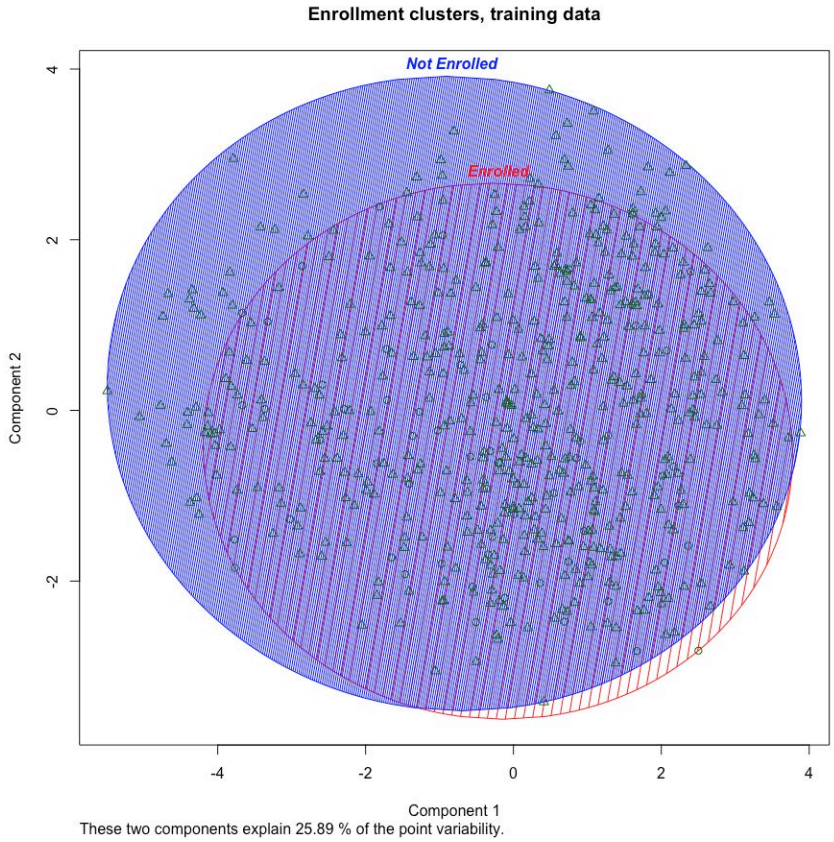

Figure 2: Cluster plot for the enrollments and non-enrollments

Five models were applied to treat the imbalanced data for enrollments and non-enrollments. The model outputs and comparisons are listed in Table 1. For instance, randomForest model correctly predicts 13 of the 28 enrollments in the holdout data (class error is 15/28=0.5357), at a cost of incorrectly predicting 55 others as enrollments who are not. Sampling the suggested ones would result in 13/68 successes (19% success rate). Adjusted Rand Index (ARI) and kappa suggest that the model is not predicting better than chance.

Model	Class Error	Success Rate	Predicted Enrollment	Adjusted Rand Index	Kappa	AUC
Logistic	0.8929	0.3000	10	0.1029	0.1166	0.6329
Logistic-SMOTE	0.5357	0.1275	102	0.0334	0.0703	0.6531
KNN	1.0000	0.0000	1	-0.0056	-0.0062	0.6975
randomForest	0.5357	0.1912	68	0.1148	0.1658	0.6705
xgboost	0.5357	0.1383	94	0.0469	0.0882	0.6189

Table 1: Table for model comparison

3.2 Identify Potential Consents

For those 872 financial well-being survey completes, 263 of them (30%) really consented. See Figure 3 as follows.

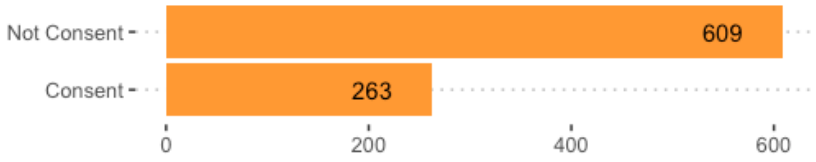

Figure 3: Plot for distribution of consent

Again, we produced a cluster plot for the consents and non-consents. Figure 4 shows that the two groups show little differentiation on the principal components, which suggests that classifying data into the groups and predicting consents could be difficult.

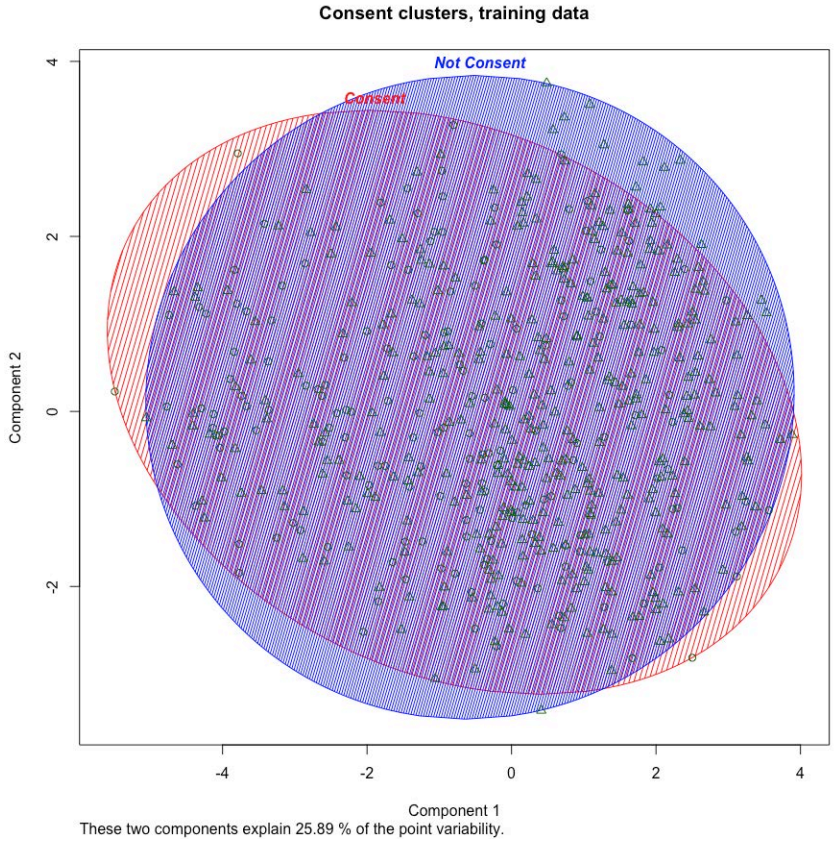

Figure 4: Cluster plot for the consents and non-consents

Similarly, five models were applied to treat the imbalanced data for consents and non-consents. The model outputs and comparisons are listed in Table 2. For instance, randomForest model correctly predicts 51 of the 91 consents in the holdout data (class error is 40/91=0.4396), at a cost of incorrectly predicting 55 others as consents who are not. Sampling the suggested ones would result in 51/106 successes (48% success rate). Adjusted Rand Index (ARI) and kappa suggest the model identifies consents in the test data modestly better than chance.

Model	Class Error	Success Rate	Predicted Consent	Adjusted Rand Index	Kappa	AUC
Logistic	0.6264	0.5667	60	0.1654	0.2865	0.7188
Logistic-SMOTE	0.3956	0.4545	121	0.1135	0.2821	0.7086
KNN	0.8462	0.5385	26	0.0733	0.1273	0.6983
randomForest	0.4396	0.4811	106	0.1390	0.3002	0.7230
xgboost	0.4286	0.4333	120	0.0914	0.2447	0.7195

Table 2: Table for model comparison

4. Summary

Standard classifier algorithms like Logistic Regression tends to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class. SMOTE significantly improves the accuracy of the classification for Logistic regression. KNN has the lowest success rate and fewest predicted events. Xgboost with balancing option performs similar to SMOTE. Balanced random forest is the best model for both enrollment and consent classification problems.

The ultimate value of each strategy depends on the cost of sampling versus the value of successful conversion. Sampling all prospects in the holdout data for enrollments would result in 28/315 successes (9% success rate), while sampling the suggested ones by random forest model would result in 13/68 successes (19% success rate).

Sampling all prospects in the holdout data for consents would result in 91/315 successes (29% success rate), while sampling the suggested ones by random forest model would result in 51/106 successes (48% success rate). Since the enrollment rate of consents is 30%, it would result in 15/106 successful enrollments (14% success rate).

Although the enrollment model has the highest success rate (19%), it has limited sample size and is not predicting better than chance. The consent model is more reliable and a better approach for model-based sampling with 14% success rate of enrollment, which is 56% higher than a random approach.

References

Breiman, L. et al. (2018). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14. https://CRAN.R-project.org/package=randomForest

Chapman, C. N., McDonnell Feit, E. (2015). R for Marketing Research and Analytics. Springer, New York.

Chen, T et al. (2020). xgboost: Extreme Gradient Boosting. R package version 1.2.0.1. https://CRAN.R-project.o rg/package=xgboost

Torgo, L. (2013). DMwR: Functions and data for "Data Mining with R". R package version 0.4.1. https://CRAN.R-project.o rg/package=DMwR

Ripley, B and Venables, W. (2020). class: Functions for Classification. R package version 7.3-17. https://CRAN.R-project.o rg/package=class