Tailored Assignment of Internet and Mail Self-Response Modes Using a Model-Based Stratification[*]

John Chesnut
U.S. Census Bureau, Washington, DC 20233

Abstract

The use of response models to group or stratify members of a target survey population into homogeneous response groups, informed by auxiliary data characterizing the sample units, is a well-established practice to reduce bias and improve reliability of survey estimates. More recently, adaptive survey design methods have extended the use of propensity models to inform tailored design changes. This paper explores the use of various Internet response models to stratify the American Community Survey (ACS) sample frame to enable a tailored initial assignment of the mail and Internet self-response modes. These models include traditional logistic regression as well as some of the more recent machine learning techniques – decision trees, random forests, boosting, support vector machines, and K-nearest neighbors. To inform the models, we augment the ACS sampling frame using administrative records. Using data from the April 2011 ACS Internet Test, we establish that offering mail or a choice of mail and Internet self-response modes are viable options for members of the low Internet response stratum.

Key Words: nonresponse, propensity models, adaptive survey design, administrative records, machine learning

1. Introduction

Mixed-mode surveys typically include an initial offering of an Internet and/or mail questionnaire self-response option as a less expensive alternative to telephone or personal interviewing. This is certainly true for the American Community Survey where all sample cases receive exclusively an Internet response option for the first three weeks of data collection. Subsequently, nonresponse cases are mailed a paper questionnaire.

However, an exclusive initial Internet offering may not be suitable for everyone. Data show that 16 percent of adults do not use the Internet (Perrin and Duggan, 2015). In addition, using ACS data, Baumgarder et al. (2014) and Nichols et al. (2014) have found evidence that portions of the population do not prefer an Internet response option. Specifically, the Baumgardner et al. study has found that with the introduction of the exclusive initial Internet offering in 2013, overall self response has decreased for households characterized as 'economically disadvantaged' or located in areas with high concentrations of minority. In addition, Nichols et al. studied responding households of the 2011 ACS Internet tests and classified them into 'hard to interview' groups (e.g., renter, minority, low education, older respondents). They observed that the exclusive offering of the Internet mode of response in lieu of a paper questionnaire negatively affected self-response among these groups. Furthermore, analysis by Roberts (2012) of data from the ACS Telephone Questionnaire

[*] Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Assistance program show an increased call volume as a result of the introduction of the Internet response option to the ACS. This was due to respondents calling in to report lack of Internet access and difficulty entering the Uniform Resource Identifier (URL) into a browser.

Given this evidence of increased respondent burden and nonresponse due to an exclusive initial Internet offering, how can we improve the mixed-mode survey design to better accommodate cases not likely to respond by Internet? A first step is to identify the cases not likely to respond via an Internet offering. In 2011, the Census Bureau conducted two national-level field tests to determine the best methodology for including an Internet option in the ACS (Tancreto et al., 2012 and Matthews et al., 2012). Using the clustering results from the Census 2010 advertising campaign research (Bates and Mulry, 2008), researchers stratified the ACS Internet Test sampling frame based on "targeted" and "not targeted" Census tracts. The researchers hypothesized and the data showed that sample cases in the targeted stratum responded via the Internet mode at a higher rate than those in the not-targeted. Our goal for this research is to refine this stratification further by using additional data sources to characterize sample cases not likely to respond by Internet prior to data collection. Given this stratification, we can then begin to tailor an initial mode assignment in such a way that we reduce respondent burden and possibly follow-up cost.

2. Methodology

2.1 Predicting Internet Response

Extending the use of the data from the April 2011 ACS Internet test, we develop our models using the sample cases from the treatment group where the contact strategy mirrors the methodology currently used in the production ACS. In addition, we link administrative record data at the address-level to supply the household-level features that will inform our models.

To build our models, we turn to the discipline of machine learning (Clark et al., 2009). Given that our data includes the Internet response outcome for each sample case, we use a supervised learning framework which entails identifying a training data set to 'learn' or train our models (machines) and then using an independent test dataset to validate the ability of our models to discriminate those not likely to respond by Internet. Following this framework, we use a 75-25 training-test split of our data. That is, we use 75 percent of the sample cases of our available data to learn our models and reserve 25 percent of the sample cases to validate or test the performance of our models in terms of prediction.

2.2 Machine Learning Models

To meet our objective of finding an adequate model for predicting Internet response, we focused on exploring various machine learning modelling techniques. We started with logistic regression as a well-known established statistical method of prediction to serve as a benchmark for comparing the performance of the other more recent modelling approaches.

We also included the popular tree-based models – classification trees, random forests, and boosting. For classification trees, we grow a single decision tree using a recursive partitioning algorithm that creates node splits iteratively from available predictors not previously used in prior node splits such that the split results in the largest decrease in impurity as measured by a Gini index or entropy measure (Briemen et al., 1984). For our random forests model, we grow many trees using bootstrapped samples of the training data. In addition, we grow trees such that we use a random sample of the available features at each node split. (Briemen, 2001, Briemen and Cutler, 2004). Similar to random forests, boosting requires growing a forest of trees. However, instead of using bootstrapped samples, the trees our grown sequentially and the training data are modified such that cases that are not predicted correctly are weighted so that they are given more emphasis in fitting subsequent trees (Schapire 1990).

In addition to the tree based models, we reviewed the use of support vector machines (Vapnik, 1998), a non-linear extension of the support vector classifier. The objective of the support vector classifier in the case where we have a two-dimensional feature space is to find the widest rectangular strip that separates the data in terms of the outcome classes.

Lastly, we reviewed the use of the simple, but powerful K-Nearest-Neighbor (KNN) technique (James et al., 2013). The objective here is to find *k* sample cases from the training data that are in close proximity to a given test sample case from the test data as measured by a distance measure (e.g., Euclidean or Manhattan distance). The predicted outcome for a given test case is based on the majority classification of its *k* neighbours sourced from the training data.

2.3 Tailored Initial Mode Assignments

After establishing a winning model for predicting Internet response, we next use our model to inform a low and high Internet response stratification. This stratification enables a tailored initial mode assignment. To do this, we use our model-derived Internet response propensities, assigned to all sample cases from the April 2011 ACS Internet Test, to stratify the cases into the low and high Internet response strata. We then evaluate the best self-response offering for those members of the low Internet response stratum by comparing the response outcome for the alternative self-response offerings of mail-only and the choice offering of mail and Internet to the exclusive initial Internet offering.

2.4 Augmenting the Sample Frame Data

The data that we used to augment the sampling frame data to supply the household-level predictor variables needed to inform our models include data from the 2010 Census, Internal Revenue Service (IRS), Info-USA, United States Postal Service (USPS), and the National Telecommunications Information Administration (NTIA). Table 1 lists these administrative record data sources and their associated variables. For more information on the record linkage results and the imputation methods used for accounting for missing data, see Chesnut (2013).

Table 1. Administrative Record Data Sources

Administrative Record Data Source	Variables
2010 Census – Housing Unit Response Data File	self-administered questionnaire, language of interview or questionnaire, proxy respondent,
2010 Census – Edited Household Data File	householder - age, race, and Hispanic origin; tenure; large household
2010 Census – Edited Person Data File	non-spousal, non-related household
2010 Census – Unedited Operation Data File	type of enumeration area, response check-in-date
Master Address File	urban-rural
Info USA	do not call flag, low-tech household
United States Post Office (USPS) – National Change of Address Database	change of address flag
National Telecommunications and Information Administration	broadband flag
Internal Revenue Service	1040 total income reported for 2010
2010 Census – Advertising (cf. Bates and Mulry, 2008)	targeted – single, detached, mobile households or advantaged homeowners

3. Results

3.1 Model Interpretation

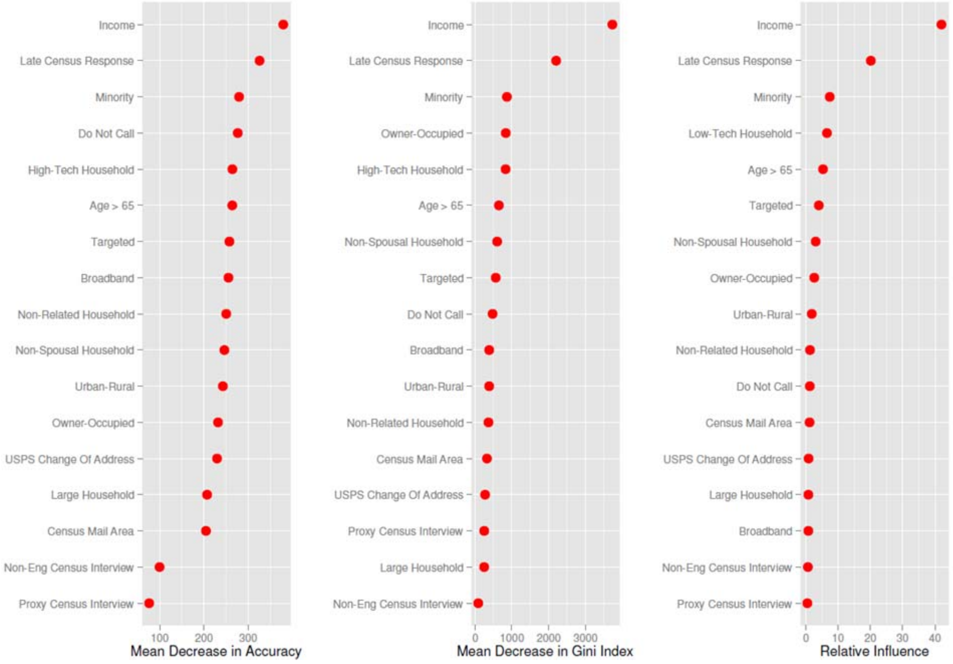
Each of the modeling methods we explored supports our goal of predicting Internet response. However, they do vary in the level of interpretability they provide in explaining how the household characteristics relate to Internet response. Logistic regression has the advantage of producing results in the form of odds ratios that help characterize those households not likely to respond via the Internet. From Table 2, our results show that households characterized as low-tech, owner-occupied, non-related, located in a 2010 Census mail enumeration area, responded via proxy in the 2010 Census, or were located in a broadband area are more likely to respond by Internet. In addition, we find that that households characterized as located in the non-targeted stratum, located in rural areas, non-spousal households, householder age greater than 65, minority, large households, late check-in of Census data for self or personal interview, non-English census interview, or households with lower income levels are less likely to respond via the Internet.

Table 2. Relating Internet Response and Household Characteristics - Odds Ratios

Variable	Odds Ratio	95% Lower Confidence Limit	95% Upper Confidence Limit
Low-tech household	1.4	1.3	1.4
Owner-occupied	1.3	1.2	1.3
Non-related	1.2	1.1	1.3
Census mail enumeration area	1.2	1.0	1.3
Census proxy response	1.2	1.1	1.4
Broadband	1.1	1.0	1.2
Non-targeted stratum (2010 Census advertising)	0.8	0.7	0.8
Rural	0.8	0.7	0.9
Non-spousal household	0.7	0.7	0.7
Age > 65	0.7	0.6	0.7
Minority	0.6	0.6	0.6
Large household	0.6	0.5	0.7
Census personal interview/late check-in	0.4	0.3	0.4
Census self-response/late check-in	0.3	0.3	0.4
Census non-English interview	0.3	0.2	0.5
Income: not reported	0.4	0.4	0.5
Income: $0-$10,000	0.6	0.5	0.7
Income: $10,001-$15,000	0.5	0.4	0.6
Income: $15,001-$25,000	0.5	0.4	0.6
Income: $25,001-$35,000	0.5	0.5	0.6
Income: $35,001-$50,000	0.6	0.6	0.7
Income: $50,001-$75,000	0.7	0.7	0.9
Income: $75,001-$200,000	0.9	0.8	1.0

From the results of our decision tree-based models that require growing forests, we can assess the importance of each of our household characteristics in informing the model using various aggregated measures across the trees in the forest. Specifically, for the random forests model, we can examine the relative importance for each of the household features in terms of their contribution to the model's accuracy in predicting an Internet response outcome. Using the 'out of bag' sample cases not included in the bootstrap samples for validation, we can derive the mean decrease in model accuracy across the trees in the forest when a given variable is removed from the model to assess the given variable's contribution. Therefore, important variables will result in larger decreases in accuracy. Figure 1 (left plot) shows the ranking of variable importance as it relates to contributing to model accuracy. From this ranking, we observe that income is the most important household feature in contributing to the model's accuracy in correctly classifying Internet response outcome followed by the 2010 Census late response indicator (379 and 326 percent respectively). The remaining features result in average decreases in accuracy ranging from 205 to 280 percent with the exception of the 2010 Census proxy status and non-English response indicator resulting in decreases of 100 and 76 percent respectively.

Figure 1. Assessing the Importance of the Household Characteristics – Accuracy, Node Purity, and Relative Influence



In addition to assessing variable importance in relation to model accuracy, we can derive for each of our household features a measure of the mean decrease in node impurity as measured by the Gini impurity index (Gini, 1912). As defined by the recursive partitioning algorithm, node splits result in smaller values of the Gini impurity index for the two children nodes compared to the parent node. Averaging the Gini decreases across all trees in the forest gives a measure of a given variable's importance in producing purer node splits. A larger decrease compared to other features indicates a greater importance in the role the feature contributes in predicting an Internet response outcome. Figure 1 (middle plot) shows the ranking of variable importance as it relates to reducing node impurity. We observe that household income and the late Census response indicator are substantially more effective than the other household features in terms of partitioning the data by Internet response outcome. In addition, we find that the household status variables for minority, owner-occupied, low-tech household, and householder age greater than 65 play an important role in partitioning the data with decreasing levels of importance for the remaining variables.

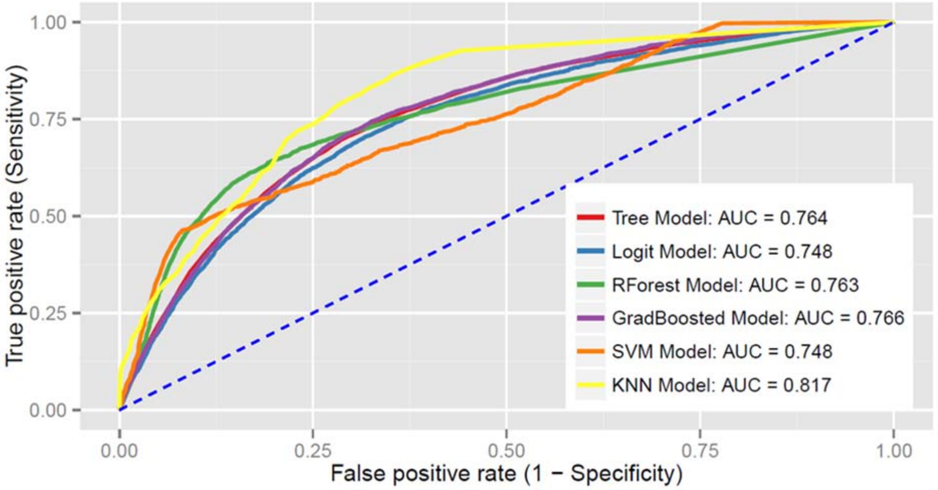
Finally, the boosting model provides a measure of the relative influence for each variable. The relative influence measure is based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman & Meulman 2003). Similar to the previous importance measures derived under the random forests model, we observe from Figure 2 (right plot) that income results in the largest relative influence followed by the late Census response indicator, From these results on variable importance, we conclude that the income characteristic is critical for predicting Internet response.

3.2 Comparing Models

To compare our modeling approaches on how well they do in predicting Internet response, we plot the Receiver Operating Characteristic (ROC) curves for each model and then calculate the area under each ROC curve (Hosmer and Lemeshow, 2000). The ROC curve plots the probability of detecting a true positive (sensitivity) and a false postive (1– specificity) for a range of possible cut-points for discriminating whether a case is an Internet response or nonresponse. The Area Under the Curve (AUC) which ranges from zero to one, provides a measure of the model's ability to discriminate the outcome of interest. Note that we used the training data set to 'learn' our models and now we use the test dataset to validate their prediction ability.

According to the AUC measures, we find that with the exception of the KNN model, all of our models demonstrate similar prediction performance. Ranking the models, logistic regression and SVM had the lowest prediction ability. Surprisingly, our single decision tree model performed just as well as the random forest model which we anticipated an improvement due to its use of bagging. Our boosting model was slightly better. Finally, our most simple model – KNN, appears to do exceptionally well. According to the AUC thresholds established by Hosmer and Lemshow (2000), the KNN model has excellent discrimination properties.

Figure 2. Comparing Models: ROC Curves

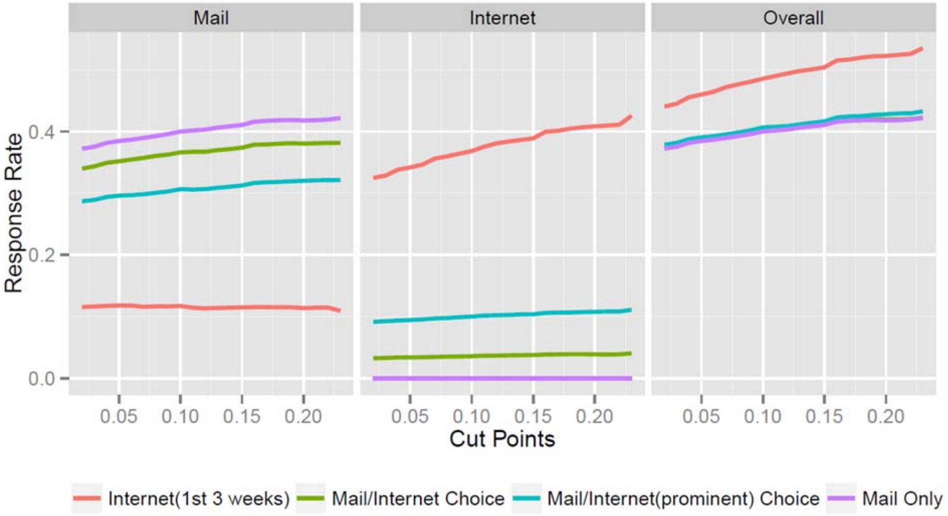


3.3 Evaluating the Tailored Initial Mode Assignments

Having established a winning model, we now stratify the April 2011 data based on our model predictions and then measure the rates of self-response within the low and high Internet response strata. Note that our model informs our stratification by associating a probability of an Internet response with each sample case. By manipulating the cut point or probability threshold for classifying a case as an Internet response or nonresponse, we can vary the strata boundaries. With our stratification, we can address the question of whether a tailored self-response offering using an alternative method would improve response among those negatively affected by initially offering exclusively an Internet response option.

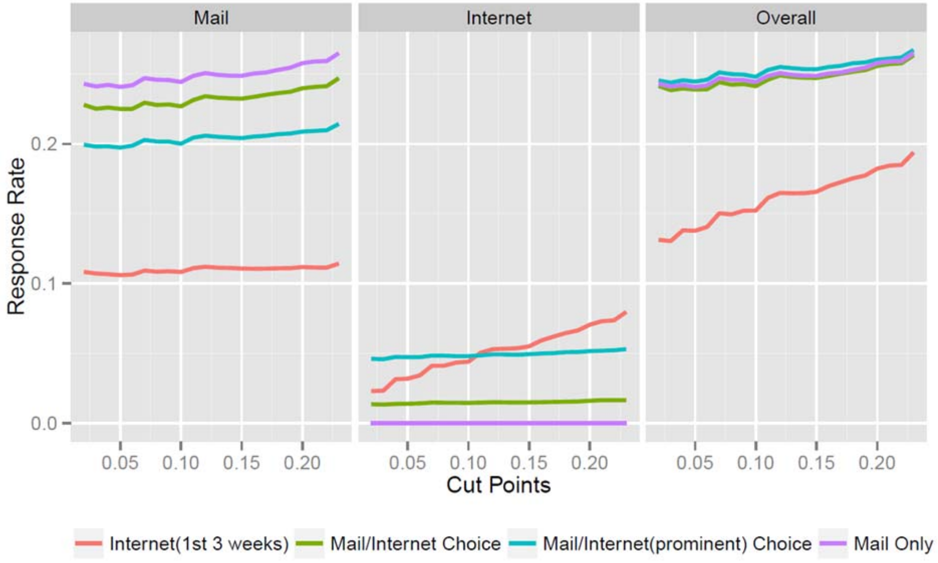
To assess the effectiveness of our stratification, we first review the graphs in Figure 3 showing the mail, Internet, and overall self response rates for the members of our high Internet response stratum for each of the treatment groups in the April 2011 Internet test over a range of cut points. Note that the choice treatment groups were initially offered both a choice of an Internet and mail response option, mail only was offered exclusively the mail response option, and Internet group was exclusively offered the Internet response option the first three weeks of data collection. We observe for the mail response rate graph (left plot), the mail only treatment leads to higher levels of mail response, followed by the treatments with a choice offering. For the Internet response rate graph (middle plot), we observe higher levels of response for the treatment with an exclusive initial Internet offering followed by the choice offerings. Finally, we observe for the overall self-response rate graph (right plot) similar levels of response for the mail-only and choice offerings, but a much higher levels of response for the Internet offering. This result demonstrates the effectiveness of our high Internet response stratification and confirms that the current ACS contact strategy is the best strategy for the high Internet response stratum.

Figure 3. High-Internet Response Stratum: Mail, Internet, and Overall Self-Response



Continuing our assessment of our stratification, Figure 4 shows the response rate results for the low Internet response stratum. Examining the mail response rate graph (left plot), we observe a similar response rate patterns observed in Figure 3. However, we note the lower levels of response for the mail-only and choice offerings. From the Internet response rate graph (middle plot), we again find similar patterns of response observed in Figure 3. However, we observe much lower levels of response for the treatment group with the exclusive initial Internet offering. This confirms that our low Internet response stratification is effectively isolating or separating out cases not likely to respond by Internet. Finally, we observe for the overall self-response rate graph (right plot), the treatment group with the exclusive Internet offering results in a much lower overall self-response compared to the alternatives. In addition to the evidence found in the literature, this provides further support that households contained in the low Internet response stratum are negatively impacted by an exclusive Internet offering. However, the overall self-response rate graph also illustrates that the mail-only or choice offerings are viable alternatives for remedying the negative impact of the exclusive Internet offering for members of the low Internet response stratum. In other words, a tailored assignment of mail-only or a choice offering would improve the level of self-response for this sub-population.

Figure 4. Low-Internet Response Stratum: Mail, Internet, and Overall Self-Response



While the previous data provide some insight into which contact strategies for the low Internet response stratum are appropriate, the question remains what is the optimal cut point for defining the strata boundaries between our low and high Internet response strata.

3.4 Optimal Cutpoint

Given that we have cost implications due to misallocating cases to our low and high Internet response strata as a result of false positive or false negatives, a logical choice for a discriminating cut point c would be one that minimizes these misclassification cost in the form of a cost or loss function.

In our case, the event of a false positive is where we classify a sample case as an Internet respondent when in fact they are not. This may incur a cost by contributing to the telephone and personal interview nonresponse follow-up workload. In addition, the event of a false negative is where we classify a sample case as a mail respondent when in fact they are willing to respond by Internet. This may incur a cost associated with the questionnaire materials, printing, mailing cost, and nonresponse.

For a given cutpoint c, we can express the full cost function as the sum of a fixed overhead cost (C_0) and the costs associated with identifying a true positive (C_{TP}), true negative (C_{TN}), false positive (C_{FP}), and false negative costs (C_{FN}), each term weighted by their probabilities of occurring (López-Ratón et al., 2014). We express these probabilities as a function of the model sensitivity (Se), specificity (Sp), and the Internet response rate (p).

$$\begin{aligned} C(c) = C_0 + C_{TP}pSe(c) + C_{TN}(1-p)Sp(c) + C_{FP}(1-p)\big(1-Sp(c)\big) \\ +C_{FN}p\big(1-Se(c)\big) \end{aligned}$$

To simplify this, we can assume a null cost associated for the occurrence of true negatives and true positives. This reduces the previous model to more simplified cost model termed the Misclassification Cost Term (Smith, 1991 and Greiner, 1995,1996).

This allows us to express our cost in terms of a cost ratio of the false negative costs to the false positive cost. Establishing an appropriate cost ratio may be easier to estimate than determining dollar amounts for the classification errors and correct classifications in the previous model. Finding the value of c at which this function is minimized will provide the optimal strata definition for our high and low Internet response strata.

$$MCT(c) = \frac{C_{FN}}{C_{FP}} p\big(1 - Se(c)\big) + (1 - p)\big(1 - Sp(c)\big)$$

4. Conclusion

Using the April 2011 ACS Internet Test data linked to administrative record data as training and test data, we were able to use machine learning to learn various modelling techniques and validate their prediction ability. As a result, we found that the tree-based and SVM models produced marginally better prediction results compared to logistic regression. However, we found that our simplest modelling approach, KNN performed exceptionally well in terms of prediction.

Using our winning model to supply the sample case-level propensities of responding via the Internet, we successfully stratified the Internet test sample into low and high Internet response strata and observed the mail, Internet, and overall self-response outcomes for the exclusive Internet offering and the alternatives – mail only, choice of mail or Internet (prominent Internet offering display and not prominent). Furthermore, we varied the strata boundaries by manipulating the propensity cut point or threshold for allocating a case to the low or high Internet response stratum.

For the high Internet response stratum, we observed similar outcomes in terms of overall self-response regardless of the type of self-response mode(s) offered, with the exception of the exclusive Internet offering which resulted in higher levels of response. This confirmed that the use of an exclusive Internet offering works well for members of the high Internet response stratum. For the low-Internet response stratum, we observed a similar pattern of response for the alternatives, however the levels of overall self-response for the Internet-only offering was consistently lower. This reinforced our motivations for conducting this study, that is portions of the population are negatively impacted by the exclusive Internet offering compared to the alternatives. In addition, our overall response rate results show that either a mail-only or choice offering are viable alternatives to the Internet-only offering for sample cases in the low-internet response stratum.

To choose an appropriate stratification boundary based on the cutpoint for classifying households as members of the low and high Internet response strata, we proposed a cost function to account for the misclassification costs attributed to allocating sample cases to the wrong stratum. Given an accurate representation of the misclassification cost, we can minimize this cost function to determine an appropriate cut point for establishing a final low and high Internet response stratification to enable a tailored self-response mode assignment reducing survey costs and respondent burden.

References

Baumgardner, S., Griffin, D. (2014). “The Effects of Adding an Internet Response Option to the American Community Survey,” 2014 American Community Survey Research and Evaluation Report Memorandum Series #ACS14-RER-21.

Bates, N. and Mulry, M. H. (2008). “Building a Segmentation Model to Target the 2010 Census Communications Campaign,” 2008 JSM Proceedings of the Section on Survey Research Methods –AAPOR paper, pp 4065-4071.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth, New York.

Breiman, L. (2001). Random forests, *Machine Learning* 45:5–32.

Breiman, L., and Cutler, A. (2004). *Random Forests*, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Chesnut, John (2013). “Model-Based Mode of Data Collection Switching from Internet to Mail in the American Community Survey,” 2013 JSM Proceedings of the Section on Survey Research Methods, pp 2209-2223.

Clarke, B., Fokoué, E., and Zhang, H. (2009). *Principles and Theory for Data Mining and Machine Learning*, New York, Springer.

Gini, C. (1912). "Italian: Variabilità e mutabilità" 'Variability and Mutability', C. Cuppini, Bologna, 156 pages. Reprinted in Memorie di Metodologica Statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).

Greiner M (1995). “Two-Graph Receiver Operating Characteristic (TG-ROC): A Microsoft Excel Template for the Selection of Cut-O Values in Diagnostic Tests." *Journal of Immunological Methods*, 185(1):145-146.

Greiner M (1996). “Two-Graph Receiver Operating Characteristic (TG-ROC): Update Version Supports Optimization of Cut-Off Values that Minimize Overall Misclassification Costs." *Journal of Immunological Methods*, 191:93-94.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression,2nd ed.*, Wiley, New York.

López-Ratón, M., Rodríguez-Álvarez, M., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). “OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests,” *Journal of Statistical Software*, 61(8).

Matthews, B., Davis, M., Tancreto, J., Zelenak, M.F., and Ruiter, M. (2012). “2011 American Community Survey Internet Tests: Results from Second Test in November 2011,” 2012 American Community Survey Research and Evaluation Report Memorandum Series #ACS12-RER-21.

Nichols, E., Horwitz, R., and Tancreto, J. (2013). “Response Mode Choice and the Hard-to-Interview in the American Community Survey.” Center for Survey Measurement Research and Methodology Research Report Series #2013-01.

Perrrin, A. and Duggan, M., (2015). “Americans’ Internet Access: 2000-2015, As Internet use Nears Saturation for Some Groups, a Look at Patterns of Adoption,” Pew Research Center, http://www.pewinternet.org/files/2015/06/2015-06-26_internet-usage-across-demographics-discover_FINAL.pdf.

Roberts, A. (2012). “An Analysis of Telephone Questionnaire Assistance Data for the Internet Notification Strategy Follow-up Test,” Census Bureau Internal Report.

Schapire, R. (1990). The strength of weak learnability, *Machine Learning*, 5(2):197–227.

Smith RD (1991). “Evaluation of Diagnostic Tests." In RD Smith (ed.), *Veterinary Clinical Epidemiology*, *3rd edition*, pp. 29-43. Butterworth-Heinemann, Stoneham, MA.

Tancreto, J., Zelenak, M. F., Davis, M., Ruiter, M., and Matthews, B. (2012), “2011 American Community Survey Internet Tests: Results from First Test in April 2011,” 2012 American Community Survey Research and Evaluation Report Memorandum Series #ACS12-RER-13.

Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.