Survey Research Methods (2023)
Vol. 17, No. 3, pp. 243-268
doi:10.18148/srm/2023.v17i3.7988
European Survey Research Association

Case Prioritization in a Panel Survey Based on Predicting Hard to Survey Households by Machine Learning Algorithms: An Experimental Study

Jonas Beste[1], Corinna Frodermann[1], Mark Trappmann[1,2], and Stefanie Unger[1]
[1]Institute for Employment Research, Nuremberg
[2]University of Bamberg, Faculty of Social Sciences, Economics and Business Administration

Panel surveys provide particularly rich data for implementing adaptive or responsive survey designs. Paradata and survey data as well as interviewer observations from all previous waves can be utilized to predict fieldwork outcomes in an ongoing wave. This manuscript contributes to the literature on how to best make use of these data in an adaptive design framework applying machine learning algorithms. In a first step, different models were trained based on past panel waves. In a second step, we assess which model best predicts fieldwork outcomes of the following wave. Finally, we apply the superior model to predict response propensities and base case prioritizations of households at risk of attrition on these predictions. An experimental design allows us to evaluate the effect of these prioritizations on response rates and on nonresponse bias. Increasing prepaid respondent incentives from 10 to 20 euros substantially decreases attrition of low propensity cases in personal as well as telephone interviews and thereby helps reduce nonresponse bias in important target variables of the panel survey.

Keywords: panel survey; adaptive design; case prioritization; machine learning

1 Introduction

Adaptive (Schouten et al., 2017; Wagner, 2008) and responsive (Groves & Heeringa, 2006) survey designs have become a standard practice during the last decades in order to achieve more balanced fieldwork outcomes in surveys and reduce bias and the variance of nonresponse weights (Peytchev et al., 2020; Wagner, 2008). While exact definitions vary, at the core of such designs, we find the use of auxiliary data to design informed interventions with the goal to affect data collection costs, data quality or both (compare Chun et al., 2017).

Adaptive or responsive designs aim at maximizing response rates and minimizing bias by specifically targeting those with a lower response probability while keeping the cost for intervention at a reasonable level (Groves & Heeringa, 2006; Tourangeau et al., 2017; Wagner, 2008). The success of responsive design strategies depends on the targeted groups' favorable reaction to a given intervention (Groves & Heeringa, 2006) as well as the availability of data including paradata such as contact information, interviewer observations or information on target variables on the sampling frame (Couper & Wagner, 2012; Schouten et al., 2013).

Contact information: Mark Trappmann, Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nürnberg, Germany (E-mail: mark.trappmann@iab.de).

Adaptive and responsive designs have focused on various mechanisms in the survey administration process to increase response rates and decrease bias focusing either on the respondent or interviewer and changing various aspects of the survey design like respondent incentives (McGonagle et al., 2022), question order (Early et al., 2017), survey mode (Calinescu & Schouten, 2015; Coffey et al., 2020) or interviewer payment (Bergmann & Scherpenzeel, 2020).

As panel attrition is increasing in panel studies around the world (Williams & Brick, 2018), the potential usefulness of adaptive designs in the context of panel surveys is increasing. Fortunately, panel surveys provide particularly rich data for implementing adaptive or responsive survey designs (Lynn, 2017; Plewis & Shlomo, 2017). Not only are data from the current wave fieldwork available, but paradata and survey data as well as interviewer observations from all previous waves can be utilized to predict fieldwork outcomes in an ongoing wave. Their early availability furthermore, allows for careful planning and modelling of predicted response propensity in advance[1]. In the recent past, this has been applied to target advance letters (Lynn, 2016), predict optimal mode (Carpenter, Burton, et al., 2018; Kaminska &

[1]Lynn rather prefers to use the term "targeted design" if variation of treatment is between subgroups (identified by such models) and not over time within a given wave of fieldwork as is the case in most adaptive and responsive designs. However, one might argue that such designs are adaptive in the sense that who gets which treatment can vary across panel waves

Lynn, 2017), or optimize timing of contact attempts (Kreuter & Müller, 2015) in a panel survey context.

Another development in the past decades is the use of machine learning algorithms in survey methodology (Buskirk, 2018; Buskirk et al., 2018; Kern et al., 2019). Previous work reveals that machine learning might be particularly useful in the context of predicting fieldwork outcomes and response propensities (Kern et al., 2019; Kern et al., 2021; Liu, 2020). While statistical methods commonly used for predicting (non)participation are usually limited to small variable sets and ignore complex patterns of interaction, machine learning algorithms are able to overcome these limitations (Zinn & Gnambs, 2022). Although there is a growing body of literature using these predictions (i.e. in weighting adjustments see Lee et al. (2010), overview in Toth and Phipps (2014)), there has been surprisingly little application to adaptive survey designs.

In a panel context Earp et al. (2012), have used response propensity predictions based on regression trees to allocate nonresponse followup funds. Early et al. (2017) use learning algorithms to order questions in an online survey in a way that maximizes survey completion.

In this article, we combine the two developments: We use machine learning algorithms in order to select panel households for prioritization in the 14th wave (2020) of the German panel study "Labour Market and Social Security" (PASS). The PASS panel survey (Trappmann et al., 2015) is a sequential mixed-mode survey of the general population that oversamples welfare benefit recipients. An adaptive survey design has until now mainly been implemented for refreshment samples (Trappmann et al., 2015).

In order to select households for prioritization, we first use data (survey data, paradata, interviewer observations) from wave 4 to 12 of the panel to train different machine learning models. In a next step we use the parameters from this training to predict wave 13 nonresponse. The quality of this prediction can be assessed and the superior model is used to finally predict wave 14 nonresponse based on data from waves 4-13.

These predictions inform an adaptive design experimentally implemented in wave 14. An experimental design allows us to investigate whether case prioritization of low propensity households in the form of targeted increased respondent incentives decreases attrition rates in these groups and reduces bias in target variables of the survey.

In the following section, we first give an overview of the PASS panel study and the variables selected to train several machine learning algorithms and finally predict response propensities (section 2.1). We then explain our approach how to evaluate performance (section 2.2) and then briefly introduce the different machine learning algorithms and their tuning parameters (section 2.3). Then we compare the algorithms on performance statistics from a test dataset (section 2.4). Based on these, we pick the optimal algorithm and describe its results (section 2.5). In the results section, we first describe the implementation of our adaptive design experiment (section 3.1), then we show effects of the adaptive design on panel retention rates by survey mode (section 3.2), before we finally simulate whether bias in target variables was reduced by the adaptive design (section 3.3). Finally, we summarize our results and discuss limitations and avenues for future research (section 4).

2 Data and Methods

2.1 Data

We implemented our experimental design in the 14th wave (2020) of the German panel study "Labour Market and Social Security" (PASS[2], Bähr et al., 2019; Trappmann et al., 2019). PASS is a representative large-scale household panel survey and one of the major German data sources for research on the labour market, unemployment and poverty dynamics. Established by the Institute for Employment Research in 2007, annual surveys with about 15.000 persons in about 10.000 households are conducted in co-operation with the fieldwork agency infas.

PASS uses a dual-frame sampling design, which combines a sample of Germany's residential population with an oversampling of households with Unemployment Benefit II receipt. Initially a household interview is carried out with the heads of all selected households. Subsequently, all members of the household aged 15 or over are interviewed. PASS uses a sequential mixed-mode design of computer-assisted face-to-face interviewing (CAPI) and computer-assisted telephone interviewing (CATI) in order to maximize response under cost-restrictions and we implemented different experimental designs in both groups (see chapter 4.1 for a detailed description of the adaptive designs). However, Wave 14 fieldwork, starting in February 2020, was severely affected by the Covid-19 pandemic. From mid-March on, CAPI interviews were stopped and respondents shifted to CATI (also see chapter 4.1 for the consequences on our experimental design strategy).

For the adaptive design, we draw on both, household as well as individual information. Our unit of analysis is the head of the household, who is first interviewed on various household-level information. After completing the household interview, the head of the household is interviewed by a person questionnaire which covers a large range of individual-level information. Our experimental design uses

[2]Data access to the Scientific Use File (SUF) is provided by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the IAB. Additional files not included in the Scientific Use File like pre-release versions of the data or contact form data (used for the predictions of outcomes) can be accessed for replication purposes

wave 4 to wave 12 for training and validation purposes of the machine learning algorithms, wave 13 for testing the performance of the different prediction models and wave 14 for the actual experiment and case prioritisation (see figure 1 and chapter 2.2 for a detailed description of the experiment). We use information from the previous waves[3] to predict the probability of dropout/re-participation in the following wave.[4] Thus, we restrict our analysis to panel cases that already participated in at least one previous wave.

Selecting appropriate machine learning algorithms that can handle a large number of variables as well as a large number of observations allows us to include a large set of variables in the training and prediction models.

To further improve the data handling, we partially coarsen the variables. We also extract aggregated information on missings, for example the total number of variables with missing values and information on missing values at variables with a relative high proportion of missing values (more than 0.5%).

We restrict our set of information on variables that were consistently asked from wave 4 to wave 13 in the household or person interviews. From the household interview we use for example the size and composition of the household, the household income and whether there are debts or residential property as well as deprivation and welfare benefit receipt. On the individual level we use sociodemographic like age and sex as well as other personal information such as education, number of friends, migration background and interview language, health and satisfaction indicators and employment status.

However, in order to select households for prioritization, we do not rely on survey data only, we also use various paradata from other data sources. First, we enrich the survey data with contact data from the previous wave. We include the total number of contact attempts and an indicator for whether an interview could be conducted.

Second, we use data of a survey answered by the interviewers, which is conducted after a person interview on interviewers' assessments of how interested respondents have been during the interview, how reliable the answers were, and whether respondents had troubles understanding the questions.

As a third source of paradata we match the interview duration of all previous interviews by using time stamps. All additional data sources have been linked to the survey data via household or personal identifier.

We end up with a final number of 74 variables to potentially use as predictors in the prediction models (see Table A1 in the appendix).

2.2 Machine learning algorithms: Training and validation approach

Machine learning algorithms can be classified as supervised or unsupervised. Supervised learning uses data where the outcome to building a machine learning model is known, while at unsupervised learning there is no supervised outcome. Machine learning algorithms can handle classification and regression problems, the former having discrete values (e.g. binary variable) as outcome and the latter real numbers (James et al., 2023, p. 15). In our study we use supervised learning for a classification problem.

To train and evaluate machine learning models the existing data is separated into sets of different use. The training set is used for learning by fitting the parameters of the model. The validation set is used to tune the parameters and the test set is used only to assess the performance of the fully-specified model. Then the fully-specified model can be applied to new data to predict the unknown outcome.

In our study, we explore different machine learning algorithms to predict the likelihood of not participating in the next wave of PASS. Our final goal is to predict the likelihood of participating for the panel cases in the gross sample of wave 14 of PASS and identify the half with the lowest probability. Therefore, we use wave 4 to 12 data to train and validate the models and test the accuracy of the predictions of the several models at previously unseen test data of wave 13. The model with the best performance on the wave 13 data is applied to wave 14 data where the outcome is yet unknown.

To evaluate how useful these algorithms are for the task at hand we use the results of a single (main effect only) logistic regression as a reference model. This procedure is commonly used to estimate the probability of dropouts in surveys (Lepkowski, 2002). To add a variable selection in advance we also perform an elastic net logit regression (eNet). Here, other machine learning algorithms can offer advantages because larger sets of variables can be considered and models and interactions between the variables do not have to be defined in advance, but are taken into account in a data-driven manner. We apply various machine learning algorithms: classification trees (CART), k-nearest-neighbour (kNN), random forest (RF) and gradient boosting machine (GBM) as well as eXtreme gradient boosting (XGB). The models are developed using information collected in previous waves (see section 2.1).

[3] We always use the information from the previous wave, except for two variables: 1) Total number of waves participated in, and 2) proportion of waves participated in since panel entry. Both variables are aggregated over all past waves.

[4] We define our outcome variable as follows: All cases with the AAPOR final disposition code of 1.1. (complete interview) for the household interview are considered as participation (coded as 0), all other final disposition codes are considered as non-participation (coded as 1).

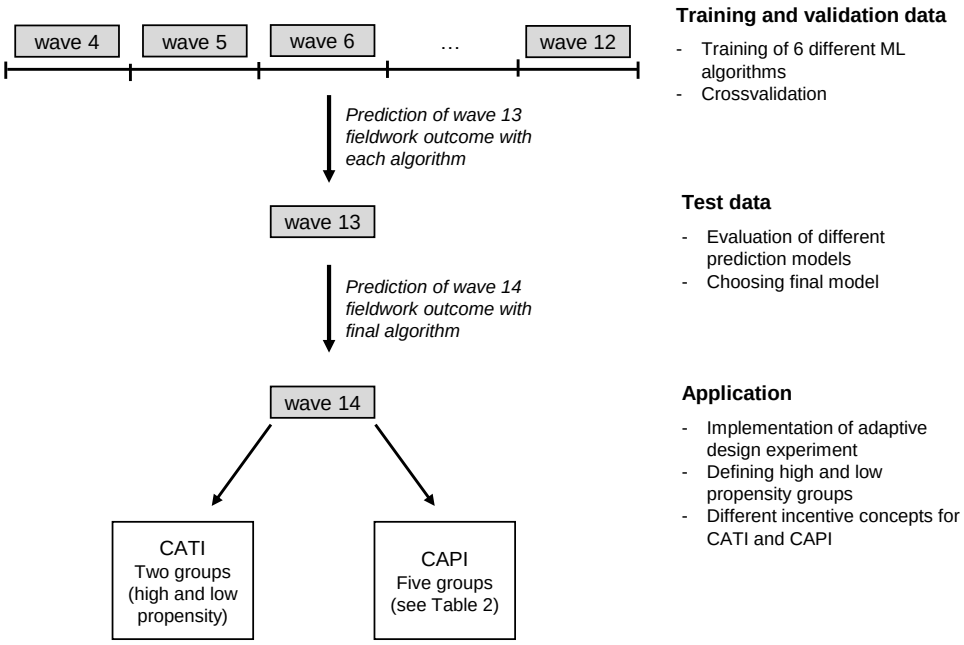

Figure 1

Timeline for training, testing and application

Before building the models, several preparations steps need to be done. All missings due to item nonresponse[5] are imputed by a single imputation using multivariate imputation via chained equations (MICE). In a pre-processing step the independent variables used in the prediction models are centered and scaled to improve modelling performance. Since our outcome (dropout in PASS) is not equally distributed (81% retention rate), we have to find a way to deal with this imbalanced data. Due to a disparity in the frequencies of the observed classes in the data, the models can become biased towards the majority class prediction (Kuhn & Johnson, 2013, p. 419). There are multiple approaches to address class imbalances. One common technique is to subsample the training data, either through up or down sampling (C. Chen & Breiman, 2004). While the former oversamples the minority class by creating duplicate obser- vations, the latter undersamples the majority class by removing data points. Neither of the two approaches is fundamentally better in correcting class imbalance. However, since additional data points are generated during upsampling, this method can result in longer computing times. Since we train multiple prediction models we decided to use downsampling.

For each algorithm, in a first step hyperparameters are tuned within the training data using k-fold cross validation to improve the estimated performance of the model. One of the most commonly used resampling methods is k-fold cross-validation. Here, training data is split randomly in k-folds. In an iteration process over all k-folds in every iteration $k-1$ folds are used to fit the model and the left-out fold to evaluate the model (Hastie et al., 2009, p. 241). Since different splits of the data can result in different results we use 3 repeated 10-fold cross-validations. In the tuning step a grid of hyperparameter settings is used, where all possible combinations of the hyperparameter are tried. The best model is chosen by evaluating which constellation of tuning parameter maximizes the cross-validated ROC AUC (area under the receiver operating characteristic curve). Finally, the performance of the final model is evaluated on the test data.

2.3 Brief introduction to the machine learning algorithms

The following section gives a brief introduction to the machine learning algorithms we use in our study.

Elastic net logistic regression

In regularization methods, model coefficients are penalized by an additional shrinkage term. This can be used for specific goals like automatic variable selection or to enhance

[5]The maximum proportion of item nonresponse is below 5%

the prediction accuracy of a model. One of the most common regularization methods is the LASSO, which penalizes the sum of absolute values of the coefficients. The higher the penalization term lambda is, the more the coefficients are shrunk toward zero. In this method some of the coefficients can become absolute zero and hence the number of relevant independent variables in the model can be reduced. In contrast, the Ridge regression penalizes the sum of the squared coefficients. Here, the coefficients are also shrunk toward zero, but never actually become absolute zero. The elastic net is a more general regularization method which uses a convex combination of the LASSO and the ridge regularization that is steered by the parameter alpha (Hastie et al., 2009, p. 661). Indeed, both methods are included as special cases. If alpha is equal to 1 the regularization is equivalent to the LASSO and if alpha is equal to 0 the regularization is equivalent to the Ridge regression. We obtained the parameters lambda and alpha by using cross-validation with a final lambda of 0.02 and an alpha of 0.4.

k-nearest neighbours

The k-nearest neighbours (*k*NN) algorithm is a non-parametric supervised machine learning algorithm that can be used to solve both classification and regression problems (James et al., 2023, p. 39). The algorithm assumes that similar data points are close to each other and is tuned by the parameter k only. This parameter defines how many neighbours are considered to predict the probability of an object. A massive advantage of this algorithm is its simplicity. It is not necessary to build a model, nor to tune other parameters or to make additional assumptions. As a down side, the algorithm is sensitive to the local structure of the data. Although this simple method is easy to implement, it works well on many problems. After 3 repeated 10-folds cross-validation the optimum value for k in our setting was estimated to be 260.

Classification Tree

Classification and regression trees (CART) can be used for supervised learning to solve either classification or regression problems (Krzywinski & Altman, 2017; Hastie et al., 2009, p. 305). A tree recursively divides the feature space (the set of values of all predictors) into regions (nodes) by searching for the best split to form the most homogenous (pure) subregions with respect to the outcome. For classifications trees with categorical outcome node purity can be measured with the Gini index. Large trees tend to overfit the training data which could lead to poor performance when they are applied to new test data. Three common ways to control the size of trees are the minimum number of observations per terminal node, the complexity parameter and pruning. The complexity parameter sets the minimum improvement in the model needed at each node. Trees are a powerful tool to discover meaningful interactions between predictors even with different scale level. Even if trees are not among the most successful learners, they are very popular because they are very intuitive and easy to interpret. Single trees are the basis for many more advance learning algorithms, like random forest and gradient tree boosting. These ensemble techniques combine multiple trees which could result in much better predictions. We decided to tune the complexity parameter cp via cross-validation and ended up with an optimum value of 0.0012.

Random Forest

Random forest is a tree-based ensemble method to resolve the overfitting problem of a single tree (Breiman, 2001). Ensemble methods combine simple models into very powerful models. Other ensemble methods are e.g. boosting or bagging. In random forest, multiple trees (mostly hundreds or thousands) are grown using different bootstrap samples of the training data and the results of all trees are combined to classify objects. To improve the performance of the learner the correlation among the trees is decreased by only using a random subset of the predictors at each split. As a downside, interpretation of multiple trees is hard to impossible (Hastie et al., 2009, p. 587). To tune the model commonly the number of grown trees, number of variables randomly sampled as candidates at each split and the minimal node size are varied. After tuning the optimum values for 500 grown trees are 10 variables of each split and a minimum of 100 observations per node.

Boosting

Boosting is an ensemble method that can be applied to many machine learning methods. In boosting models are built sequentially by using the information from the previous models (Berk, 2006). Here, we use two boosting methods: Gradient Boosting Machine (Friedman, 2001) and eXtreme Gradient Boosting (T. Chen & Guestrin, 2016). Similar to random forest both methods base on single trees. In contrast to random forest, the trees in boosting are not independent of each other, but build on each other. A central parameter is the shrinkage parameter lambda which controls at which rate boosting learns. Beside this, the tree complexity and the number of boosting iterations can be tuned in both methods. We also use other algorithm specific tuning parameters.[6]

Table A2 in the appendix gives an overview of the machine learning algorithms used and their tuning parameters.

[6]While in Gradient Boosting Machine we also use the minimum node size as a tuning parameter, in eXtreme Gradient Boosting we additionally use the minimum loss reduction, the subsample ratio of columns, the minimum sum of instance weight and subsample percentage.

2.4 Model evaluation

To evaluate the performance of the prediction models on the test data of wave 13 we use the ROC AUC score. This statistic looks at the trade-off between the true positive rate and the false positive rate. This is equivalent to calculating the rank correlation between predictions and targets. This is advantageous when a good ranking of predictions is of high importance. All performance statistics are displayed in Table 1.

When we compare the results of the several models the main effects logistic regression model serves as the baseline benchmark. Here, the ROC AUC score is 0.6836, the elastic net logistic model ROC AUC score is 0.6837. The CART model has the lowest value, followed by the kNN model. In comparison to the complex ensemble techniques, the logistic model performs quite well. The random forest model has the highest ROC AUC score with 0.6863, followed by the GBM model.

We also use the balanced accuracy rate (percentage of correctly classified objects in both classes with equal balanced weight), the sensitivity rate (true positive rate: the rate of correctly classified non-respondents) and the specificity rate (true negative rate: the rate of correctly classified respondents). To calculate these statistics a threshold at which value of the predicted probability an object is treated as a respondent respectively as a non-respondent has to be chosen. An optimal threshold with respect to balance between false positive and true positive rates can be computed.[7] Finally, for the threshold at the median we use the precision (positive predictive value: number of true positives divided by the total number of positive predictions). This measure can be used as an indication of the rate at which incentives are targeted efficiently.

Regarding the performance statistics at the optimal threshold we prefer models with a good balance at all indicators rather than those with only high sensitivity or specificity. So, our main focus is on the balanced accuracy. For all of the statistics, the higher the numbers, the better the model predicts the outcomes. At the optimal threshold for the logistic regression model the balanced accuracy rate is 0.6403, the sensitivity rate is 0.5347 and the specificity rate is 0.7459. The accuracy of the model is mostly gained by its high specificity. Since we are interested in correctly identifying the non-respondents, this model is not optimal for our purposes. For the elastic net logistic model the balanced accuracy rate is 0.6417, the sensitivity rate is 0.6506 and the specificity rate is 0.6328. Again, the CART and kNN models perform worst, even so they have a better sensitivity rate than the logistic regression model. Comparing the three ensemble methods, GBM has the highest specificity rate (0.6650), XGB the highest sensitivity rate (0.6751) and random forest the highest balanced accuracy (0.6454).

Beside the performance metrics at the optimal threshold, due to our study specific design to give additional incentives to the panel cases in the lower half of the participating probability and our goal to correctly identify the half with the lowest participation probability, we want to know how many of the observed non-respondents of wave 13 are assigned to the low propensity cases. Therefore, we also calculate the rates with threshold at the median value (which creates equal sized groups of predicted respondents and non-respondents). Here, we are interested mostly in the rate of correctly identified non-respondents (sensitivity rate). A higher number indicates a better prediction model for our purposes. For the median value, we also show the precision.

Looking at the performance statistics at the threshold at the median, we are mostly interested in the sensitivity rate, since this matches the design of the experiment. For the logistic model, balanced accuracy rate is 0.6284, sensitivity rate is 0.7081 and specificity rate is 0.5487. For the elastic net logistic model, balanced accuracy rate is 0.6253, sensitivity rate is 0.7030 and specificity rate is 0.5475. The XGB model has the highest balanced accuracy (0.6352) and the highest sensitivity rate (0.7191). The random forest model performs similar with a balanced accuracy of 0.6341 and a sensitivity rate of 0.7174. The models with the highest are the XGB (0.2726) and the random forest (0.2720).

2.5 Choosing the final model

After comparing the performance statistics of all models on the test data we have to choose one model to work with in the following steps. Even if there is not a clear winner in all aspects, looking at the difference statistics with slightly better performance in ROC AUC and the balanced accuracy at the optimal threshold and with similar values at the threshold at median, we decided to use the random forest prediction model. We rebuild the random forest model with all data of wave 4 to 13 using the selected tuning parameters to estimate the participation likelihood of the panel cases in the gross sample of wave 14.

In random forest it is possible to get the variable importance of the predictors in use (Hastie et al., 2009, p. 539) and partial dependence plots (Hastie et al., 2009, p. 369) to further investigate the results (see Figures A1 and A2 in the appendix). Applying the random forest model with the final parameters to the test data the number of contact attempts in the previous wave has the biggest impact. With more contact attempts in the previous wave higher predicted probabilities of attrition (PPA) occur. Additionally, the age of the respondent as well as the duration of the last household and person interview and the number of previous waves are

[7] This can be shown in ROC curve plots. In this type of plot the optimal threshold would be a value on the curve that is closest to the top-left of the plot with the maximum sum of true-positive and false-negative values.

Table 1

Performance statistics in test data

		At optimal threshold			At threshold at median			
	ROC AUC	Balanced accuracy rate	Sensitivity rate	Specificity rate	Balanced accuracy rate	Sensitivity rate	Specificity rate	Precision
Logit	0.6836	0.6403	0.5347	0.7459	0.6284	0.7081	0.5487	0.2684
eNet	0.6837	0.6417	0.6506	0.6328	0.6253	0.7030	0.5475	0.2665
kNN	0.6545	0.6142	0.5981	0.6302	0.6040	0.6658	0.5421	0.2591
CART	0.6413	0.6210	0.6277	0.6142	0.6128	0.6684	0.5572	0.2609
RF	0.6863	0.6454	0.6430	0.6478	0.6341	0.7174	0.5509	0.2720
GBM	0.6833	0.6422	0.6193	0.6650	0.6315	0.7132	0.5499	0.2704
XGB	0.6819	0.6400	0.6751	0.6049	0.6352	0.7191	0.5512	0.2726

important. Higher ages and higher numbers of participated waves occur with lower PPA while longer durations occur with higher PPA. Furthermore, variables on the financial and material situation have a high importance. Higher household net incomes as well as higher household savings and higher satisfaction with housing occur with lower PPA. For material deprivation, we see a u-shaped relationship with the PPA. Interviewer characteristics like age and experience of the interviewer also have an impact. While the PPA stay relatively constant over the age till about 60 years, at this point the PPA increase strongly. Anyway, the PPA decrease with higher experience of an interviewer. Additionally, the PPA decrease with higher number of friends and higher satisfaction with health. Finally, the relationship between the social integration and the PAA is u-shaped.

3 Results

3.1 Implementation of the adaptive design experiments

Based on the models described above, we implemented an experimental design in wave 14 (in the year 2020) of PASS fieldwork. Only households that had responded in the previous wave (no temporary dropouts, no refreshment cases) were subject to the experiment. We randomized separately for addresses initially issued to face-to-face (CAPI) and to telephone (CATI) mode. Randomization—like response propensity estimation—was performed at the level of households.

First, the panel sample was divided in half at the median of the predicted response propensities. We denote the upper 50% of the predicted response propensities as "high propensity" and the bottom half as "low propensity" cases. Since, on average, CAPI cases have a somewhat higher response propensity, this leads to an unequal distribution of the two categories across modes: 1421 out of 2306 (62%) CATI cases were in the low propensity group compared to 1845 out of 4287 (43%) CAPI cases.

All previous wave respondents in high propensity households in both modes received the general prepaid respondent incentive of 10 euros per person mailed in advance and interviewers received their usual payment for these cases.

For low propensity cases assigned to CATI mode, the experimental design was straightforward and consisted of only two groups. Those 50% (710 cases) who were randomized to receiving no preferential treatment were assigned to the same incentive conditions as high propensity cases (885 cases). The other 50% (711 cases) were assigned to a doubled prepaid respondent incentive of 20 euros per person.

For low propensity cases assigned to CAPI mode, a more complex design was chosen. Not only were respondent incentives manipulated in the same way as described for CATI. In addition, and orthogonal to the respondent incentive experiment, interviewers received an extra premium for a random half of the low propensity cases. However, receiving an extra incentive for part of their workload might lead interviewers to substitute effort from cases without extra remuneration to cases with these extra payments. To control for this, we selected about 20% of the interviewers to a condition where they never received extra remuneration for any of their cases. This leads to five experimental groups among low propensity cases assigned to CAPI, each consisting of roughly one fifth of this subsample: 1) Cases without any extra respondent or interviewer incentives within interviewers that were randomly selected to never receive extra incentives, 2) Cases without any extra respondent or interviewer incentives within interviewers with a mixed assignment, 3) Cases with an extra interviewer incentive only, 4) Cases with an extra respondent incentive only, 5) Cases with both, an extra interviewer and respondent incentive. Tables 2 gives an overview of the experimental groups and their sizes in CAPI mode.

Due to this experimental design with unequal treatment probabilities by mode, we will differentiate all results by ini-

Table 2

Experimental groups and their sizes—CAPI mode

	Percentages			Absolute group sizes		
Respondent Incentives:	Yes	No	Total	Yes	No	Total
Interviewer incentive: Yes	18	19	37	342	360	702
Interviewer incentive: No	19	19	38	356	362	718
Interviewer incentive: Never	22		22	425		425

Without Interviewer identification: 3% (N=54), High Propensities: N=2442

tially assigned mode. Note, that by design this initial mode can be switched during the course of the fieldwork whenever target households cannot be contacted or request a mode switch. Respondent incentives are not affected by mode switches. Interviewer incentives are only paid in CAPI.

Note also, that the fieldwork of the 2020 panel wave in which the experiment took place was severely affected by Covid-19 containment measures. Fieldwork began on February 14th, 2020. On March 16th, 2020 all CAPI interviews were stopped and all cases with an available telephone number were shifted to CATI. Until then 1559 households had been completed in CAPI and 878 in CATI. From then on until the end of the fieldwork only telephone interviews were conducted. These were mainly conducted by CATI interviewers, although towards the end of the fieldwork CAPI interviewers became involved in telephone interviews as well. Due to these specific circumstances, we refrain from giving detailed results for the experiments involving interviewer incentives. We instead focus our analyses on the effects of respondent incentives that were unaffected by the circumstances.

Before we turn to the results of our experiments we demonstrate that our models were able to effectively predict cases with different response propensities in wave 14. In Table 3, we show response rates by predicted decile of the response propensity distribution. The first column labelled "no incentive" contains only cases without extra incentives and can thus be compared across the whole distribution. For CATI, we find a strictly monotonous trend until including the eighth decile, with response rates in wave 14 rising from 50% in the lowest decile of the predicted response propensity distribution to 93% in the eighth decile. In the ninth and tenth decile values remain close to this maximum. For CAPI we observe a similar pattern with wave 14 response rates rising from 30% in the lowest decile to 86% in the highest (with a reversal of the strictly monotonous trend only between deciles 6 and 7).

These results demonstrate that we were able to validly predict response propensities and to do so particularly well in the bottom half of the response propensity distribution.

We return to this table later in order to discuss the effectiveness of incentives across response propensity strata.

3.2 Effects of case prioritization by mode

For completeness sake we will first display results for all experiments including the interviewer incentive experiments that could only be effective in the first six weeks of the fieldwork. We will however, focus our interpretation on the outcome of the respondent incentives experiment.

Tables 4 shows main effects of the increased incentive for low propensity cases in CATI. While the response rate for high propensity cases is 91%, it is much lower for the low propensity cases. Among the low propensity cases, those treated with the doubled respondent incentive show a 7 percentage point higher response rate (69%) than those who received the regular respondent incentive (62%). This difference is statistically significant at the 5% level.

Table 5 shows slightly higher response rates for the two groups with increased respondent incentives (66%) compared to the three groups with regular respondent incentives (62, 64, and 60%). The overall effect of respondent incentives on response rates is 5 percentage points and statistically significant at the 5% level. The interviewer incentive has no effect, which was to be expected due to the early ending of CAPI fieldwork.

We conclude that doubling incentives for low propensity cases can be an effective strategy in order to raise response rates among this group. As we only targeted low propensity cases, we cannot, however, derive whether this additional incentive is more effective for them than it would be for high propensity cases. We can, however, investigate whether the effect differs across response propensity strata in the lower half of the distribution. This result is contained in Table 3. This Table shows no clear trend. For CATI the effect is almost of the same size in the lowest decile (7 percentage points) as in the third decile (6) while it is largest in the fifth (12) and smallest in the second decile (2).

A similar pattern emerges for CAPI. Again, the first (3) and fifth (3) decile show almost no differences and considerably larger effects can be found in the second (7) and fourth (7) decile. Thus, while the doubled respondent incentive effectively increases response rates in the whole lower half of the response propensity distribution, our data show no evi-

Table 3

Response rates by predicted decile of the response propensity distribution - CATI and CAPI

	CATI				CAPI			
Respondent incentive:	No		Yes		No		Yes	
Decile	%	*N*	%	*N*	%	*N*	%	*N*
1	50	271	57	263	30	71	33	52
2	62	156	64	149	45	220	52	127
3	67	98	73	104	60	291	65	158
4	74	102	83	99	68	276	75	175
5	77	83	89	96	75	289	78	186
6	88	156		-	82	510		-
7	89	137		-	76	528		-
8	93	135		-	82	531		-
9	90	172		-	84	493		-
10	93	285		-	86	380		-

Table 4

Main effects of increased incentive for low propensity cases in CATI

Treatment	Mean	*N*
High Propensities	90.85	885
Low—with respondent incentive	68.78	711
Low—without respondent incentive	61.54	710
Total	75.02	2306

dence that this is particularly the case for households with extremely low response propensity.

3.3 Simulation of effects of prioritization on bias

Adaptive designs usually target groups that are otherwise underrepresented in order to achieve a more balanced outcome of different groups in the sample and thereby reduce nonresponse bias.

We try to answer the question whether our adaptive design reduced nonresponse bias with respect to important target variables. To this end we first try to reconstruct the counterfactual situation what the outcome would have been without the adaptive design elements.

We will then assess a second counterfactual situation, namely what would have happened, had we applied the adaptive design (extra respondent incentives) to all low propensity cases. Our basic assumption is straightforward: We assume that without prioritization the incentivized cases in the low propensity half would have produced the nonresponse bias we observed for the non-incentivized cases in the low propensity half in our sample.

We explain our approach based on the data in Table 6. This Table refers to household income and to CATI mode.

Starting with the first row, we see that in the complete CATI sample the mean household income measured in the wave before (Wave 13) was 2182 euros. Among those that responded to Wave 14, the average Wave 13 household income was 2347 euros. Thus, there is a nonresponse bias of substantial size (165 euros) at a response rate of 75%. The second row shows how these numbers change if we look only at the high propensity half of the sample. Here, the initial sample had a mean household income of 2931 euros, while the mean household income of those who responded is only marginally larger at 2951 euros at a response rate of 90%. This indicates that nonresponse bias is only a minor issue among those with high predicted response propensity. In row three, we look only at low propensity cases that were randomized to the group that received no additional incentive and that reflects a fieldwork without an adaptive design. The average Wave 13 household income in that part of the sample was 1758 euros. Note, that this is considerably less than in the high propensity half which shows us how strongly related response propensity is to household income. Among those who participated from this part of the sample, the average wave 13 household income was 1649 euros. Again, this points to a positive bias of 108 euros at a response rate of 62%. Now, for comparison, we turn to the low propensity cases that were randomized to be specifically incentivized. The Wave 13 average income in this part of the sample is 1773 euros. Note, that any baseline differences to the former group occur by chance only (and are not significant at the 5%-level, though in this case the difference is quite large and significant at the 10%-level). Among those who responded

Table 5

Main effects of increased incentive for low propensity cases in CAPI

Treatment	Mean	*N*	Mean	*N*
High Propensity	81.86	2442		
Low—respondent and interviewer incentive	66.37	342	66.33	698
Low—respondent incentive only	66.29	356		
Low—interviewer incentive only	61.94	360		
Low—neither respondent nor interviewer incentive	63.81	362	61.03	1147
Low—never	57.88	425		
Total	73.76	4287	63.03	1845

Table 6

Stepwise simulation of hypothetical outcomes under no prioritization (row 5) and under prioritization of all low propensity cases (row 6)

		Mean household income					
		Gross sample wave 14 €	Realized sample wave 14 €	Nonresponse bias €	Gross sample N	Net sample N	Response rate %
1.	Total sample	2182	2347	165	2306	1730	75
2.	High propensity cases only	2931	2951	20	885	804	91
3.	Non-incentivized low propensity cases only	1649	1758	108	710	437	62
4.	Incentivized low propensity cases only	1773	1868	95	711	489	69
5.	Hypothetical outcomes if no one had been incentivized	2182	2362	179	-	-	-
6.	Hypothetical outcome if all had been incentivized	2182	2323	141	-	-	-

from that part of the sample, the average household income is 1868 euros. The bias is at 95 euros at a response rate of 69%. This nicely demonstrates the two components of the adaptive design that ideally work together to decrease bias. In the low propensity group that received additional incentives nonresponse bias (95 euros) is somewhat lower than in the low propensity group that received no additional incentives (108 euros). But what is more important is that the incentive brings more of these low propensity cases (69% vs. 62%) into the final mix which helps reduce the bias that would otherwise be present.

We have chosen an example with a rather large chance baseline difference (1649 vs 1773 euros; significant at the 10%-level) in order to motivate why it is not adequate to simply compare average income among respondents for both groups (1758 vs. 1868).

To estimate the composite effect of these two mechanisms, we will construct the following counterfactual situations:

For the hypothetical case that no one would have been prioritized we add the result in the high propensity half to that of the non-prioritized low propensity half. Finally, we assume that the prioritized low propensity half would have produced the same bias and response rate as the non-prioritized low propensity half. Weighing all these hypothetical outcomes by their respective gross sample sizes and expected realized sample sizes (given the group specific response rate), we get equation (1) for the hypothetical outcome if no case was prioritized:

$$\begin{aligned}\tilde{\bar{x}}_{\text{NI}} = \Big(&\bar{x}_{\text{High}} \cdot n_{\text{High}} \\ &+ (\bar{x}_{\text{Low, NI}} - \bar{X}_{\text{Low, NI}}) \cdot n_{\text{Low, NI}} \\ &+ \left(\bar{X}_{\text{Low., I}} + (\bar{x}_{\text{Low, NI}} - \bar{X}_{\text{Low, NI}})\right) \\ &\cdot N_{\text{Low, I}} \cdot \frac{n_{\text{Low, NI}}}{N_{\text{Low, I}}}\Big) \\ &\div \left(n_{\text{High}} + n_{\text{Low, NI}} + \frac{N_{\text{Low, I}} \cdot n_{\text{Low, NI}}}{N_{\text{Low, NI}}}\right) \quad ,\end{aligned} \tag{1}$$

with upper case variables refer to the gross samples, lower case variables refer to the realized sample, and "I" and "NI" refer to icentivized and not icentivized samples. "High" and "Low" stand for high and low propensity cases. Thus:

n_{High} Number of realized high propensity cases;

N_{High} Size of gross sample of high propensity cases;

$n_{\text{Low, NI}}$ Number of realized non-incentivized low propensity cases;

$N_{\text{Low, NI}}$ Size of gross sample of non-incentivized low propensity cases;

$n_{\text{Low,I}}$ Number of realized incentivized low propensity cases;

$N_{\text{Low, I}}$ Size of gross sample of incentivized low propensity cases;

$\bar{x}_{\text{High}}$ Mean of variable x in high propensity realized sample;

$\bar{X}_{\text{High}}$ Mean of variable x in high propensity gross sample;

$\bar{x}_{\text{Low, NI}}$ Mean of variable x in non-incentivized low propensity realized sample;

$\bar{X}_{\text{Low, NI}}$ Mean of variable x in non-incentivized low propensity gross sample;

$\bar{x}_{\text{Low,I}}$ Mean of variable x in incentivized low propensity realized sample;

$\bar{X}_{\text{Low, I}}$ Mean of variable x in incentivized low propensity gross sample.

For the hypothetical case that all cases in the low propensity half would have been prioritized, we add the result in the high propensity half to that of the prioritized low propensity half. Finally, we assume that the non-prioritized low propensity half would have produced the same bias and response rate as the prioritized low propensity half, leading to equation (2)

$$\begin{aligned}\tilde{X}_I = &\Big((\bar{x}_{\text{High}} \cdot n_{\text{High}}) + (\bar{x}_{\text{Low,I}} - \bar{X}_{\text{Low, I}}) \cdot n_{\text{Low,i}}\Big) \\ &+ \left(\left(\bar{X}_{\text{Low, NI}} + (\tilde{\bar{x}}_{\text{I}} - \bar{X})\right) \cdot \frac{N_{\text{Low, NI}} \cdot n_{\text{Low,I}}}{n_{\text{Low, I}}}\right) \\ &\div \left(n_{\text{High}} + n_{\text{Low, I}} + \frac{N_{\text{Low, NI}} \cdot n_{Low,I}}{N_{\text{Low, I}}}\right)\end{aligned} \tag{2}$$

Our fieldwork that used prioritization in only a random half of all low propensity cases resulted in a nonresponse bias of 165 euros for household income. Had we applied no prioritization it would have been at 179 euros. While prioritizing all low propensity cases would lead to a bias of 140 euros.

Tables A3a to A3g in the appendix contain corresponding results and simulation for CATI other target variables, while Tables A4a to A4g contain results for the same variables for CAPI.

Apart from the one example we described in detail, we limit ourselves to summarizing the effects of case prioritization on simulated bias by mode for a set of variables that were measured in the previous wave and that have been found in the past to be candidates for nonresponse bias: These variables are household income, household size, age, having been born in Germany, satisfaction with the living standard, number of close friends, self-rated social inclusion and health satisfaction.

For most of these variables, household income, household size, age, having been born in Germany, satisfaction with living standard, we find significant differences between respondents and non-respondents before case prioritization in both modes. In addition, we find significant differences between respondents and non-respondents before prioritization for self-rated social inclusion for CATI.

Out of these variables, we find in CATI that a full case prioritization would have the potential to decrease bias for household income (-22%), age (-28%), satisfaction with the living standard (-20%), social inclusion (-47%), and being born in Germany (-19%). On the downside, bias is increased for household size (+56%).

For CAPI, we find that a full case prioritization would have the potential to decrease bias for age (-25%), household size (-79%), and being born in Germany (-29%). Bias is increased for household income (+2%) and satisfaction with the living standard (+64%).

In summary, case prioritization reduces bias for eight variables, while it unintentionally increases bias for three variables.

4 Summary and Discussion

We have demonstrated in this article how machine learning algorithms can be used to inform adaptive survey design in the context of a panel survey. Training different algorithms on twelve prior waves of data collection and using wave 13 fieldwork outcomes as test data, we identified Random Forest (with 500 trees, minimum 100 objects in each node and 10 randomly selected variables at each split) as the algorithm that performed best at predicting future fieldwork outcomes of households from survey and paradata of past panel waves.

In wave 14 of the mixed-mode (CATI/CAPI) PASS panel survey, we implemented an adaptive design experimentally.

About half of the households estimated to be low propensity, were extra incentivized with a 20 euros prepaid incentive instead of 10 euros. We could demonstrate that this incentive increased response rates for low propensity cases significantly by more than 7% in CATI and more than 5% in CAPI. We find no systematic differences of this effect size within the low propensity half between the cases with the lowest response rate and cases with response rates close to the median.

We could show that by shifting response rates of households with a low predicted response propensity upwards we are able to moderately reduce nonresponse bias with respect to target variables of the survey like household income, age, social inclusion and being born in Germany.

While these results are encouraging and the survey managers of PASS have decided to utilize the adaptive design without the experimental evaluation in future waves of PASS, there remain some limitations and open questions for future research.

Several limitations apply to this research. The onset of the Covid-19-pandemic in Germany and the first strict lockdown fell into the second month of fieldwork and had a strong influence on the CAPI-fieldwork in the 2020 PASS wave including switches to telephone interviews for all cases that had not been interviewed within the first month of fieldwork. This limits the generalizability of the findings that are based on outcomes from that wave.

The overall retention rate for cases with previous wave interviews was at an all time low in wave 14, although we consider it quite acceptable at 74% given the extreme disruptions. Furthermore, while panel retention used to be larger in CAPI than CATI in previous waves, it was the other way around in wave 14 due to the unrequested mode-switches from CAPI to CATI after the first month. However, while overall turnout might have been different, especially for CAPI cases, the experimental and control groups are affected in the same way by all these disruptions.

Unfortunately, we also could not evaluate the second experiment in which we promised interviewer premiums for successful CAPI interviewers in low propensity households due to the early ending of CAPI fieldwork.

Another limitation is that—as our focus was on prioritization of low propensity households in an adaptive design framework—we cannot evaluate whether the extra incentives would have been as helpful in increasing response rates in the high propensity half as in the low propensity half. It seems likely however, that there would have been ceiling effects that would at least for CATI have prohibited increases of 7% from a baseline of 91%.

One should also note that model selection was based on the quality of predictions for one specific wave (Wave 13). The selected parameters that were optimal in the pre-pandemic world, might not have been an optimal choice for data collection during a pandemic. In future research, this might be addressed by including a temporal cross-validation in the model selection which uses multiple time points for model evaluation and model selection.

A promising avenue for future research would be a more sophisticated adaptive design in which response propensities are estimated based on past waves but then updated in the light of paradata from the ongoing wave fieldwork (Schouten et al., 2018). This could be a useful strategy to specifically target those groups with increased incentives that though their predicted response propensity was initially high show indications of a lower expected outcome in the paradata of the ongoing fieldwork. Given findings that events between waves can cause panel attrition (Trappmann et al., 2015), households who exhibit such a drop in expected response propensity might be likely candidates for households with changes with respect to important substantial variables (e.g. (un)employment, family status), that might be of special importance to keep estimates of change unbiased.

We also left the question for future research if other strategies than increased respondent incentives—like interviewer premiums, call schedules, targeted information—could successfully be used instead of or in combination with increased financial incentives to better fulfil the important task of keeping panel cases with a relatively low predicted response propensity in a long-running panel.

References

Bähr, S., Beste, J., Coban, M., Dummert, S., Friedrich, M., Frodermann, C., Gundert, S., Müller, B., Schwarz, S., Teichler, N., Trappmann, M., Unger, S., Wenzig, C., Berg, M., Cramer, R., Dickmann, C., Gilberg, R., Jesske, B., & Kleudgen, M. (2019). Panel Study Labour Market and Social Security (PASS) – Version 0618 v1 [Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB)]. https://doi.org/10.5164/IAB.PASS-SUF0618.de.en.v1

Bergmann, M., & Scherpenzeel, A. (2020). Using field monitoring strategies to improve panel sample representativeness: Application during data collection in the Survey of Health, Ageing and Retirement in Europe (SHARE) [Special issue on Fieldword Monitoring Strategies for Interviewer-Administered Surveys]. *Survey Methods: Insights from the Field (SMIF)*. ttps://surveyinsights.org/?p=12720

Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, *34*(3), 263–295.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buskirk, T. D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Survey Practice*, *11*(1), 1–13.

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, *11*(1), 1–10.

Calinescu, M., & Schouten, B. (2015). Adaptive survey designs to minimize survey mode effects—a case study on the Dutch Labor Force Survey. *Survey Methodology*, *41*(2), 403–426.

Carpenter, H., Burton, J., et al. (2018). Adaptive push-to-web: Experiments in a household panel study. *Understanding Society (Working Paper 2018-05)*.

Chen, C., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. University of California.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

Chun, A. Y., Barry, S., & James, W. (2017). -editorial. *Journal of Official Statistics*, *33*(3), 571–577.

Coffey, S., Reist, B., & Miller, P. V. (2020). Interventions on-call: Dynamic adaptive design in the 2015 national survey of college graduates. *Journal of Survey Statistics and Methodology*, *8*(4), 726–747. https://doi.org/10.1093/jssam/smz026

Couper, M. P., & Wagner, J. (2012). Using paradata and responsive design to manage survey nonresponse.

Early, K., Mankoff, J., & Fienberg, S. E. (2017). Dynamic question ordering in online surveys. *Journal of Official Statistics*, *33*(3), 625–657. https://doi.org/10.1515/jos-2017-0030

Earp, M., Mitchell, M., McCarthy, J. S., & Kreuter, F. (2012). Who is responsible for the bias? Using proxy data and tree modeling to identify likely nonrespondents and reduce bias. *Proceedings of the Fourth International Conference on Establishment Surveys*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *169*(3), 439–457.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Datamining, inference, and prediction*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in R* (2nd. edtion). Springer.

Kaminska, O., & Lynn, P. (2017). The implications of alternative allocation criteria in adaptive design for panel surveys. *Journal of Official Statistics*, *33*(3), 781–789.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93.

Kern, C., Weiß, B., & Kolb, J.-P. (2021). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. *Journal of Survey Statistics and Methodology*, smab009.

Kreuter, F., & Müller, G. (2015). A note on improving process efficiency in panel surveys with paradata. *Field Methods*, *27*(1), 55–65.

Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature Methods*, *14*(8), 757–758.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346.

Lepkowski, M. P., J. M.and Couper. (2002). Nonresponse in longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 259–272). Wiley.

Liu, M. (2020). Big data meets survey science: A collection of innovative methods. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Using machine learning models to predict attrition in a survey panel* (pp. 415–433). Wiley.

Lynn, P. (2016). Targeted appeals for participation in letters to panel survey members. *Public Opinion Quarterly*, *80*(3), 771–782. https://doi.org/10.1093/poq/nfw024

Lynn, P. (2017). From standardised to targeted survey procedures for tackling non-response and attrition. *Survey Research Methods*, *11*(1), 93–103.

McGonagle, K. A., Sastry, N., & Freedman, V. A. (2022). The effects of a targeted "early bird" incentive strategy on response rates, fieldwork effort, and costs in a national panel study. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smab042

Peytchev, A., Pratt, D., & Duprey, M. (2020). Responsive and adaptive survey design: Use of bias propensity during data collection to reduce nonresponse bias. *Journal of Survey Statistics and Methodology*, *10*(1), 131–148.

Plewis, I., & Shlomo, N. (2017). Using response propensity models to improve the quality of response data in longitudinal studies. *Journal of Official Statistics*, *33*(3), 753–779.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, *39*(1), 29–58.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., & Wagner, J. (2018). A Bayesian analysis of design parameters in survey data collection. *Journal of Survey Statistics and Methodology*, *6*(4), 431–464.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. CRC Press.

Toth, D., & Phipps, P. (2014). Regression tree models for analyzing survey response. *Proceedings of the government statistics section*, 339–351.

Tourangeau, R., Michael Brick, J., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(1), 203–223.

Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., Schwarz, S., Teichler, N., Unger, S., & Wenzig, C. (2019). Data resource profile: Panel study labour market and social security (PASS). *International Journal of Epidemiology*, *48*(5), 1411–1411g.

Trappmann, M., Gramlich, T., & Mosthaf, A. (2015). The effect of events between waves on panel attrition. *Survey Research Methods*, *9*(1), 31–43.

Wagner, J. R. (2008). *Adaptive survey design to reduce nonresponse bias* (Doctoral dissertation). University of Michigan. Ann Arbor, USA.

Williams, D., & Brick, J. M. (2018). Trends in US face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186–211.

Zinn, S., & Gnambs, T. (2022). Analyzing nonresponse in longitudinal surveys using Bayesian Additive Regression Trees: A nonparametric event history analysis. *Social Science Computer Review*, *40*(3), 678–699.

Appendix A
Tables

Table A1

Variables potentially used as predictors in the prediction models

Label	Scaling	Data source
Mode of household interview	Categorical	Household interview
Language of household interview	Categorical	Household interview
Federal state, generated	Categorical	Household interview
Type of renting?	Categorical	Household interview
Size of household	Metric	Household interview
Deprivation index	Metric	Household interview
Receipt social security benefit/pension supplement for old age	Binary	Household interview
Receipt care allowance	Binary	Household interview
Receipt payments from other person	Binary	Household interview
Household is paying other person	Binary	Household interview
Household income (in EUR)	Metric	Household interview
Savings of HH (in EUR)	Categorical	Household interview
Total amount of debts (in EUR)	Categorical	Household interview
Income from letting and leasing	Binary	Household interview
Other income from estate	Binary	Household interview
Child under age 4 in HH	Binary	Household interview
Child under age 15 in HH	Binary	Household interview
Current Unemployment Benefit II receipt in HH	Binary	Household interview
Move since prewave	Binary	Household interview
Sex of interviewee	Categorical	Person interview
Age	Metric	Person interview
Member of a religious community	Binary	Person interview
Marital status	Categorical	Person interview
Satisfaction with health	Metric	Person interview
Satisfaction with housing	Metric	Person interview
Satisfaction with standard of living	Metric	Person interview
Social integration	Metric	Person interview
Social position: Top-bottom-scale	Metric	Person interview
Satisfaction with one's life in general	Metric	Person interview
Student at school or university/apprentice?	Categorical	Person interview
Employed: Mini-job (marginal employment)?	Binary	Person interview
Receipt payments from statutory pension insurance?	Binary	Person interview
Receipt of private/company pension?	Binary	Person interview
Close friends/family members outside household	Binary	Person interview
Number of close friends/family members outside household	Metric	Person interview
Actively engaged in: Union	Binary	Person interview
Actively engaged in: Political party	Binary	Person interview
Actively engaged in: Church community	Binary	Person interview
Actively engaged in: Clubs such as music/sport/culture clubs	Binary	Person interview
Actively engaged in: Another organization	Binary	Person interview

Continues on next page

Continued from last page

Label	Scaling	Data source
Number of doctor's visit, last three months	Metric	Person interview
Officially recognised disabilities?	Categorical	Person interview
Other serious health restrictions	Binary	Person interview
Type of health insurance	Categorical	Person interview
Indicator: Provide care for relatives/friends on regular basis?	Binary	Person interview
Born in Germany?	Binary	Person interview
Own child under 18 years in household	Binary	Person interview
Highest school qualification	Categorical	Person interview
Highest vocational qualification	Categorical	Person interview
Current occupation (>450 EUR)	Binary	Person interview
Current unemployment	Binary	Person interview
(Un)married/ registered partner in household?	Binary	Person interview
Record linkage consent	Binary	Person interview
Duration interview (household questionnaire)	Metric	Household interview
Duration interview (person questionnaire)	Metric	Person interview
Total number of waves participated in	Metric	Household register
Proportion of waves participated in since panel entry	Metric	Household register
Sample affiliation	Categorical	Contact data
Number of contacts until household interview was realized	Metric	Contact data
Total number of variables with missing values	Metric	Household/Person interview
Missing: Deprivation - From medical insurance not reimbursed treatments?	Binary	Household interview
Missing: Deprivation - Pay rent on time?	Binary	Household interview
Missing: Household income (in EUR)	Binary	Household interview
Missing: Savings of household (in EUR)	Binary	Household interview
Missing: Total amount of debts (in EUR)	Binary	Household interview
Missing: Social position - Top-bottom-scale	Binary	Person interview
How interesting was the interview for the respondent?	Binary	Interviewer survey
How good did the respondent understand the questions all in all?	Binary	Interviewer survey
How reliable appear the answers of the respondent all in all?	Binary	Interviewer survey
Any difficulties in answering certain questions?	Binary	Interviewer survey
Interviewer: Sex	Binary	Interviewer survey
Interviewer: Work experience as an interviewer in years	Metric	Interviewer survey
Interviewer: Highest school-leaving certificate	Categorical	Interviewer survey
Interviewer: Age	Metric	Interviewer survey

Table A2

Machine learning algorithms and tuning parameters used

Algorithm	Tuning parameter	Tuning grid
Elastic Net logistic regression (eNet)	Amount of regularization (lambda)	0(0.01)0.2
	Weight of L1 and L2 penalties (alpha)	0(0.1)1
k-Nearest-Neighbours (kNN)	Number of neighbours considered (k)	20(20)400
Classification and regression tree (CART)	Complexity Parameter (cp)	0.0008(0.0001)0.02
Random Forest (RF)	Number of variables to possibly split at in each node	6, 8, 10, 12, 14, 16, 20, 30
	Minimal node size	30, 50, 100, 150, 200, 300, 500
Gradient Boosting Machine (GBM)	Number of Boosting Iterations	200, 300, 400, 500, 600
	Complexity of the tree	1, 2, 3, 4
	Learning rate (Shrinkage)	0.02, 0.04, 0.06, 0.08
	Minimal Node Size	100, 150, 200, 300, 400
eXtreme Gradient Boosting (XGB)	Number of Boosting Iterations	100, 200, 300
	Complexity of the tree	1, 2, 3
	Learning rate (Shrinkage)	0.02, 0.05, 0.1
	Minimum Loss Reduction	0, 0.5, 1
	Subsample Ratio of Columns	0.4, 0.6, 0.8
	Minimum Sum of Instance Weight	1, 2
	Subsample Percentage	0.8, 1

Table A3

Stepwise simulation of hypothetical outcome for variable household size under no prioritization (row 5) and under prioritization of all low propensity cases (row 6)—CATI

		1.	2.	3.	4.	5.	6.
	Household size	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	2.35	2.28	−0.063	2306	1730	0.75
2.	High propensity cases only	2.00	1.98	−0.016	885	804	0.91
3.	Non-incentivized low propensity cases only	2.52	2.54	0.023	710	437	0.62
4.	Incentivized low propensity cases only	2.61	2.56	−0.058	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	2.35	2.30	−0.050			
6.	Hypothetical outcome if all had been incentivized	2.35	2.27	−0.078			

Table A4

Stepwise simulation of hypothetical outcome for variable age under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Age	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	43.79	45.60	1.810	2306	1730	0.75
2.	High propensity cases only	52.69	52.66	−0.030	885	804	0.91
3.	Non-incentivized low propensity cases only	38.07	39.52	1.448	710	437	0.62
4.	Incentivized low propensity cases only	38.42	39.43	1.003	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	43.79	45.90	2.115			
6.	Hypothetical outcome if all had been incentivized	43.79	45.30	1.513			

Table A5

Stepwise simulation of hypothetical outcome for variable born in Germany under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Born in Germany	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	0.63	0.70	0.073	2306	1730	0.75
2.	High propensity cases only	0.95	0.95	−0.002	885	804	0.91
3.	Non-incentivized low propensity cases only	0.42	0.48	0.062	710	437	0.62
4.	Incentivized low propensity cases only	0.44	0.50	0.057	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	0.63	0.71	0.081			
6.	Hypothetical outcome if all had been incentivized	0.63	0.70	0.065			

Table A6

Stepwise simulation of hypothetical outcome for variable satisfaction with the living standard under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Satisfaction with the living standard	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.70	6.86	0.165	2306	1730	0.75
2.	High propensity cases only	7.19	7.21	0.014	885	804	0.91
3.	Non-incentivized low propensity cases only	6.30	6.49	0.187	710	437	0.62
4.	Incentivized low propensity cases only	6.48	6.63	0.154	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	6.70	6.88	0.180			
6.	Hypothetical outcome if all had been incentivized	6.70	6.84	0.145			

Table A7

Stepwise simulation of hypothetical outcome for variable number of close friends under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Number of close friends	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	7.76	7.86	0.098	2306	1730	0.75
2.	High propensity cases only	7.39	7.47	0.081	885	804	0.91
3.	Non-incentivized low propensity cases only	8.19	8.28	0.098	710	437	0.62
4.	Incentivized low propensity cases only	7.80	8.12	0.319	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	7.76	7.79	0.033			
6.	Hypothetical outcome if all had been incentivized	7.76	7.93	0.172			

Table A8

Stepwise simulation of hypothetical outcome for variable self-rated social inclusion under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Self-rated social inclusion	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.99	7.11	0.116	2306	1730	0.75
2.	High propensity cases only	7.27	7.25	−0.019	885	804	0.91
3.	Non-incentivized low propensity cases only	6.82	7.05	0.232	710	437	0.62
4.	Incentivized low propensity cases only	6.83	6.94	0.110	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	6.99	7.15	0.153			
6.	Hypothetical outcome if all had been incentivized	6.99	7.07	0.080			

Table A9

Stepwise simulation of hypothetical outcome for variable health satisfaction under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CATI

		1.	2.	3.	4.	5.	6.
	Health satisfaction	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.82	6.76	−0.067	2306	1730	0.75
2.	High propensity cases only	6.38	6.40	0.025	885	804	0.91
3.	Non-incentivized low propensity cases only	7.12	7.22	0.102	710	437	0.62
4.	Incentivized low propensity cases only	7.09	6.93	−0.160	711	489	0.69
5.	Hypothetical outcomes if no one had been incentivized	6.82	6.82	−0.004			
6.	Hypothetical outcome if all had been incentivized	6.82	6.70	−0.125			

Table A10

Stepwise simulation of hypothetical outcome for variable household size under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Household size	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	2.17	2.13	−0.046	4287	3162	0.74
2.	High propensity cases only	2.10	2.11	0.010	2442	1999	0.82
3.	Non-incentivized low propensity cases only	2.30	2.14	−0.163	1147	700	0.61
4.	Incentivized low propensity cases only	2.23	2.20	−0.028	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	2.17	2.11	−0.065			
6.	Hypothetical outcome if all had been incentivized	2.17	2.16	−0.014			

Table A11

Stepwise simulation of hypothetical outcome for variable age under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Age	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	43.89	44.83	0.940	4287	3162	0.74
2.	High propensity cases only	49.45	49.35	−0.107	2442	1999	0.82
3.	Non-incentivized low propensity cases only	36.60	37.16	0.567	1147	700	0.61
4.	Incentivized low propensity cases only	36.40	36.90	0.507	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	43.89	44.93	1.042			
6.	Hypothetical outcome if all had been incentivized	43.89	44.67	0.781			

Table A12

Stepwise simulation of hypothetical outcome for variable born in Germany under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Born in Germany	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	0.78	0.83	0.043	4287	3162	0.74
2.	High propensity cases only	0.93	0.92	−0.001	2442	1999	0.82
3.	Non-incentivized low propensity cases only	0.59	0.66	0.073	1147	700	0.61
4.	Incentivized low propensity cases only	0.60	0.65	0.049	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	0.78	0.83	0.049			
6.	Hypothetical outcome if all had been incentivized	0.78	0.82	0.035			

Table A13

Stepwise simulation of hypothetical outcome for variable satisfaction with the living standard under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Satisfaction with the living standard	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.83	6.92	0.090	4287	3162	0.74
2.	High propensity cases only	7.06	7.11	0.049	2442	1999	0.82
3.	Non-incentivized low propensity cases only	6.59	6.61	0.016	1147	700	0.61
4.	Incentivized low propensity cases only	6.44	6.60	0.167	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	6.83	6.90	0.073			
6.	Hypothetical outcome if all had been incentivized	6.83	6.95	0.120			

Table A14

Stepwise simulation of hypothetical outcome for variable number of close friends under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Number of close friends	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.91	6.80	−0.101	4287	3162	0.74
2.	High propensity cases only	6.88	6.84	−0.040	2442	1999	0.82
3.	Non-incentivized low propensity cases only	6.97	6.64	−0.324	1147	700	0.61
4.	Incentivized low propensity cases only	6.89	6.89	0.003	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	6.91	6.76	0.147			
6.	Hypothetical outcome if all had been incentivized	6.91	6.88	−0.027			

Table A15

Stepwise simulation of hypothetical outcome for variable self-rated social inclusion under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Self-rated social inclusion	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.94	6.99	0.048	4287	3162	0.74
2.	High propensity cases only	7.09	7.08	−0.008	2442	1999	0.82
3.	Non-incentivized low propensity cases only	6.76	6.79	0.038	1147	700	0.61
4.	Incentivized low propensity cases only	6.73	6.90	0.173	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	6.94	6.97	0.030			
6.	Hypothetical outcome if all had been incentivized	6.94	7.02	0.076			

Table A16

Stepwise simulation of hypothetical outcome for variable health satisfaction under no prioritization (row 5) and under prioritization of all low propensity cases (row 6) - CAPI

		1.	2.	3.	4.	5.	6.
	Health satisfaction	Share/mean in gross sample wave 14	Share/mean in realized sample wave 14	Nonresponse bias in wave 14	Number of observations in gross sample	Number of observations in net sample	Response rate
1.	Total sample	6.61	6.55	−0.057	4287	3162	0.74
2.	High propensity cases only	6.32	6.35	0.030	2442	1999	0.82
3.	Non-incentivized low propensity cases only	7.03	6.93	−0.096	1147	700	0.61
4.	Incentivized low propensity cases only	6.92	6.83	−0.087	698	463	0.66
5.	Hypothetical outcomes if no one had been incentivized	6.61	6.55	−0.062			
6.	Hypothetical outcome if all had been incentivized	6.61	6.56	−0.048			

Appendix B
Figures

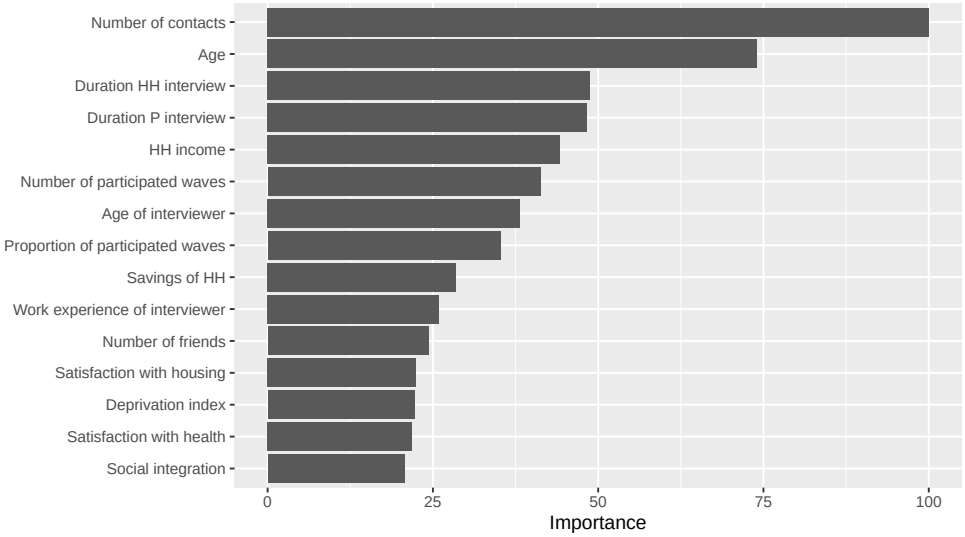

Figure B1

Variable importance of the 15 most important variables

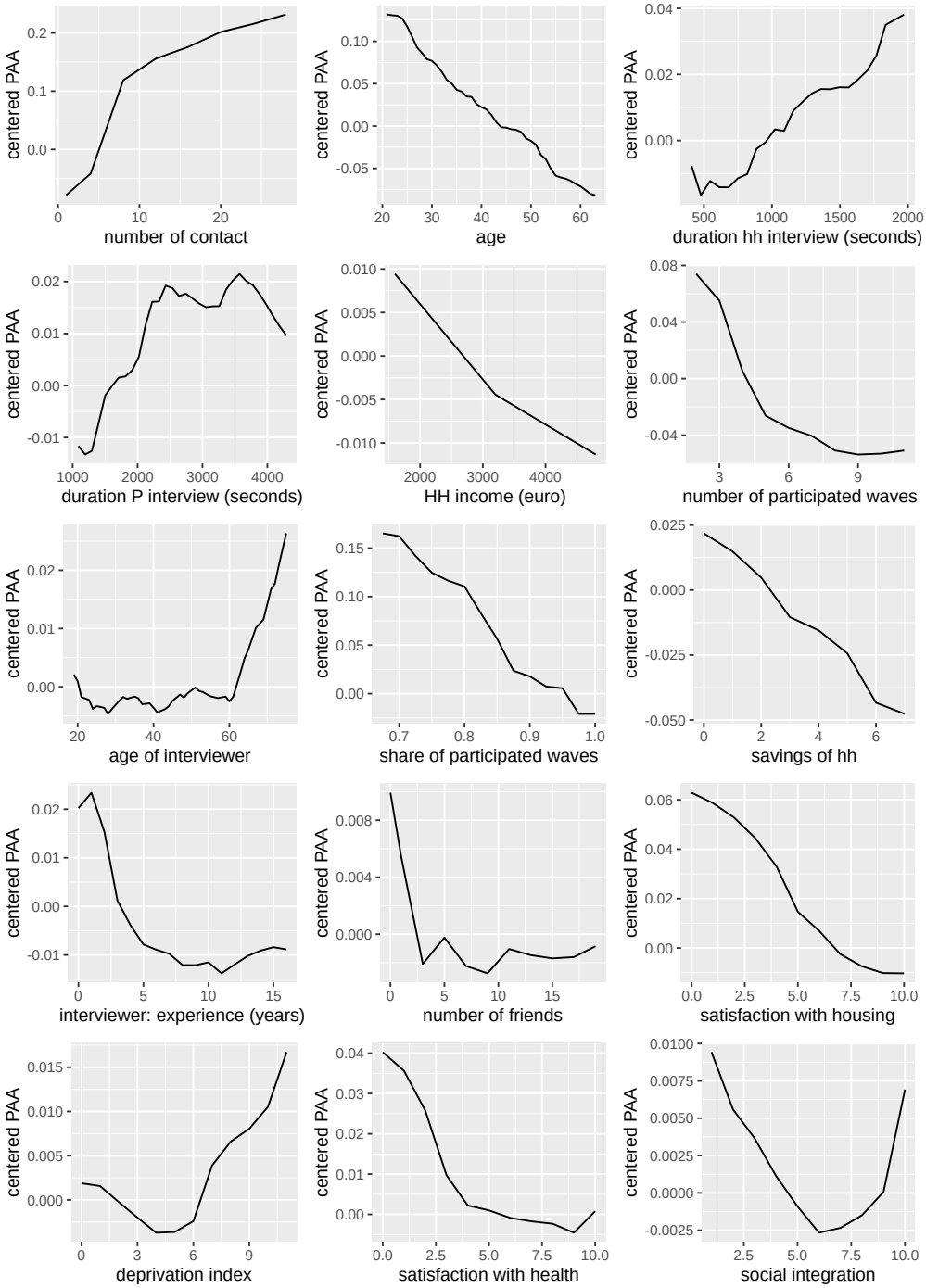

Figure B2

Partial dependence plots for important variables