*5th International Conference on Advanced Research Methods and Analytics (CARMA2023)*
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16416

# Use of machine learning techniques in non-probabilistic samples

**Jorge Rueda[1], Beatriz Cobo[2], Luis Castro[2]**
[1]Department of Statistics and Operations Research, University of Granada, Spain, [2]Department of Quantitative Methods for Economics and Business, University of Granada, Spain.

***Abstract***

*Non-probabilistic surveys are increasingly used because they are easy and cheap to carry out. Even official statistical agencies are starting to use this type of surveys in their research, due to the difficulty and the amount of resources needed to carry out probabilistic surveys, which are currently the best option due to their reliability. When non-probabilistic surveys are used, the classical estimation methods cannot be used since the initial conditions for carrying them out are not met, so over the years new estimation techniques have been emerging in this type of sampling. Some of the most relevant estimation techniques currently being used are those related to machine learning techniques.*

*In this work we focus on the estimation technique for non-probabilistic samples statistical matching, which can be enhanced and improved if we complement it with a machine learning technique known as XGBoost. We are going to study a variable of interest extracted from a real non-probabilistic survey carried out during the COVID-19 pandemic, and check if by applying such estimations we obtain better results than without applying this type of techniques.*

***Keywords:*** *Machine learning; non-probabilistic sampling; statistical matching; XGBoost.*

## 1. Introduction

The major strength of probability sampling is that the probability selection mechanism permits the development of statistical theory to examine the properties of sample estimators. The weakness of all nonprobability methods is that no such theoretical development is possible; as a consequence, nonprobability samples can be assessed only by subjective valuation (Kalton, 1983). Over the years the development of non-probabilistic surveys has boomed and many techniques have been developed to calculate reliable estimates from non-probabilistic survey data.

Many advanced in artificial intelligence models, such as deep learning techniques, have shown remarkable accuracy in prediction. Artificial intelligence models perform poorly when dealing with relatively small data sets, while machine learning models have good predictive performance on smaller data sets. However, a single machine learning approach often leads to overfitting and difficulty handling the large number of imbalanced data sets that occur in real world problems. To make up for the shortcomings of a single machine learning method, the conjoint learning technique based on the GBDT (Gradient Boost Decision Tree) algorithm was developed and has gradually become the mainstream approach in the field of learning research automatic. eXtreme Gradient Boosting (XGBoost) is a highly efficient booster set learning model originated from the decision tree model, which uses the tree classifier for better prediction results and higher operational efficiency.

This technique has been used in many settings, for example Li and Yao (2018) classify gene mutations using machine learning models, XGBoost and SVM, in the hope of improving gene mutation classification performance. In terms of performance of the two qualifying models, XGBoost outperformed SVM. From the confounding metrics, it could be seen that XGBoost had better predictive ability, especially for those with enough classes featured. Liu et al. (2021) used a mortality prediction model using the XGBoost decision tree model for patients with acute kidney injury in the intensive care unit, and compared its performance with that of three other machine learning models, logistic regression (LR), support vector machines (SVM), and random forest (RF) being XGBoost the best performing algorithm in this study. Castro-Martin et al. (2021) test the potential of the XGBoost algorithm in the most important estimation methods that integrate data from a probability survey and a non-probability survey. At the same time, a comparison is made of the effectiveness of these methods for the elimination of biases. The results show that the proposed estimators based on gradient increasing frameworks can improve the representativeness of the survey with respect to other classical prediction methods. The proposed methodology is also used to analyze a real sample from a non-probabilistic survey on the social effects of COVID-19. Cui et al. (2022) created an accurate prediction model using machine learning techniques, such as logistic regression, XGBoosting machine,

random forest, neural network, gradient boosting machine, and decision tree, to predict 3-month mortality specifically among lung cancer patients with bone metastases according to readily available clinical data. Today, people tend to use credit cards for their payment efficiency, but credit cards also provide a new opportunity for fraud. Companies and researchers have been trying to come up with a method to tell if a transaction is fraudulent. Cai and He (2022) propose a hybrid model based on the combination of TabNet and XGBoost. A dataset provided by IEEE-CIS is used in this investigation, which contains many records of transactions and whether they are fraudulent.

Our work focuses on the combination of data obtained through probabilistic and non-probabilistic surveys with the aim of obtaining more reliable estimates through XGBoost. As a non-probabilistic survey, we will base ourselves on the survey carried out by Pérez et al. (2020) and as a probabilistic survey the CIS Barometer of May 2020.

## 2. Methodology

Let U be the finite population of interest of size N, $s_v$ a non-probabilistic (or volunteer) sample of size $n_v$, from which we measure a vector of auxiliary variables $x = (x_1, \dots, x_p)$ and the variable of interest $y$ that we want to know about the population U. Normally the results we obtain from this kind of samples present different types of biases, especially the one known as selection bias, which appears if there is a significant difference between the individuals in our sample and those not sampled. To correct this type of bias there are several techniques, which depend on the type of auxiliary information available (Rueda et al., 2020). If we have a reference probability sample $s_r$, of which we only know the same vector of auxiliary variables as in $s_v$, we can apply the technique known as statistical matching, based on superpopulation models.

### *2.1. Statistical Matching (SM)*

Also known as Mass Imputation, it was developed by Rivers (2007). It is based on modeling the relationship between the variable of interest and the vector of auxiliary variables, using the non-probabilistic sample $s_v$ to predict the values of the variable of interest in the probabilistic sample $s_r$, since they are unknown. Assuming that the population of interest U is a realization of a superpopulation model $m$:

$$y_i = m(x_i) + e_i, \;\; i = 1, \dots, N$$

Where $m(x_i) = E_m[y_i|x_i]$ y $e \sim N(0, \sigma)$. That is, we can model the relationship between the variable of interest and the auxiliary variables using some model (which we will call SM). From such a model we estimate the prediction of the values of $y$ in the probability sample $s_r$, using the values of the auxiliary variables in that sample, of the form:

$$\hat{y}_i = E_{SM}[y_i|x_i, 1_i], \;\; i \in s_r$$

$1_i$ will have a value equal to one if the i-th individual belongs to the probability sample $s_r$, and will be zero when it does not belong to this sample. Depending on the model we use, we will have different expressions of $\hat{y}_i$. Once we obtain the prediction of our variable of interest *y*, we can construct the estimator of our choice in the form (case of the estimator of the population total):

$$\hat{Y}_{SM} = \sum_{i \in s_r} \hat{y}_i \cdot w_{ri}$$

Being $w_{ri}$ the design weight for the i-th element of the reference sample. We see that in this technique the most important step is to predict the variable of interest *y*, to perform this step we can use machine learning techniques that produce a prediction as accurate as possible. In our work we will use the technique known as XGBoost, which is producing excellent results both in the prediction of variables and in the estimation of inclusion probabilities for non-probabilistic samples.

### *2.2. XGBoost Estimator*

In our case we will use the XGBoost technique to obtain the predicted values of the response variable for the probabilistic sample $s_r$. This machine learning technique works as a group of decision trees, which establish branches (different paths) as a function of $x_i$ until a final value $\hat{y}_i$ is obtained (Chen and Guestrin, 2016). The expression of $\hat{y}$ using XGBoost is:

$$\hat{y}_i^{XG} = \phi(x_i) = \sum_{k=1}^{K} f_k\,(x_i), \;\; f_k \in F$$

where K is the number of decision trees, $F = \{f(x) = \omega_{q(x)}\}$ with $q: \mathbb{R}^p \to T$ representing the structure of each tree, and $\omega_i$ is the score of the i-th final node. Finally we obtain the predicted value $\hat{y}_i^{XG}$ by summing the scores of each tree, which are designed to minimise the following objective function:

$$L(\phi) = \sum_i l\,(\hat{y}_i^{XG}, y_i) + \sum_k \Omega\,(f_k)$$

where $l$ is a function that measures the error in the estimates, and which must be differentiable and convex (i.e. difference squared). To regularise this function, there is a $\Omega(f)$ that penalises trees with too many final nodes *T* and exaggeratedly high scores $\omega$, of the form:

$$\Omega(f) = \gamma T + \frac{\lambda ||\omega||^2}{2}$$

being $\gamma$ and $\lambda$ hyperparameters that directly influence the regularisation of the function. This regularisation serves to control the so-called overfitting, which appears when the machine learning model has a behaviour specific to the type of data we train it with, producing bad results when the input data are different to those we have used to train the model (Hawkins, 2004). Because of this it is very important what values these hyperparameters have, that can be taken arbitrarily or by hyperparameter optimization (i.e. by cross validation). Finally $L(\phi)$ is minimised with the gradient tree boosting method, developed by Friedman (2001). This allows us to converge to the minimum value of a function through an iterative process (gradient descent), training the models by giving more importance to the data for which previous models have failed (boosting). To improve its performance, XGBoost also implements other techniques such as shrinkage, to limit the influence of each individual tree, among others (Chen and Guestrin, 2016).

Once we estimate the values of the variable of interest for the individuals of the probability sample $s_r$ by XGBoost $\hat{y}_i^{XG}$, we obtain that the estimator of the population total using statistical matching is:

$$\hat{Y}_{SM}^{XG} = \sum_{i \in s_r} \hat{y}_i^{XG} \cdot w_{ri}$$

## 3. Application

Combining statistical matching with XGBoost as the chosen machine learning method is a relatively costly process which, in addition, has to be repeated for each variable of interest. In this case, we have chosen the following variable from the survey conducted by Pérez et al. (2022) during the Spanish lockdown caused by the COVID-19 pandemic: "Would you be willing to continue teleworking after the lockdown?". We could then evaluate the interest of the population in working remotely now that, even though it is not mandatory anymore, it has emerged as an interesting option.

The percentage of individuals responding affirmatively considering only the non-probabilistic sample in a naive way would be 26.2%. However, it is preferable to consider possible biases caused by the snowball methodology used during the distribution of the online survey. For this reason, we also consider the CIS Barometer of May 2020. The variables in common between our non-probabilistic sample and the auxiliary probabilistic sample are the following: state, province, urban density, sex, age, education level, employment status, last electoral vote, intended electoral vote and confidence in the government during the pandemic.

Once the bias reduction process is completed, we find out that the percentage of individuals who would not mind continuing to telework is actually 33.1% instead of the initial 26.2%. Therefore, we observe a significant increase from the initial impression before considering a more advanced analysis.

## 4. Conclusions

In this work, we have considered a method combining statistical concepts and advanced machine learning techniques in order to improve the reliability of the estimations for a variable of interest. We have also observed, via a real application, how relevant applying said method can be for the final conclusions obtained.

When a strict methodology is not considered for carrying out a survey, it is important to consider these kinds of methods in order to avoid possible biased results.

## References

Cai Q. & He J. (2022). Credit Payment Fraud detection model based on TabNet and Xgboot. 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, pp. 823-826, doi: 10.1109/ICCECE54139.2022.9712842.

Castro-Martín, L., Rueda, M.M., Ferri-García, R., & Hernando-Tamayo, C. (2021). On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics* 9, 23: 2991. https://doi.org/10.3390/math9232991.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cui, Y., Shi, X., Wang, S., Qin, Y., Wang, B., Che, X., & Lei, M. (2022). Machine learning approaches for prediction of early death among lung cancer patients with bone metastases using routine clinical characteristics: An analysis of 19,887 patients. *Frontiers in Public Health*, 10. https://doi.org/10.3389/fpubh.2022.1019168.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Hawkins, D. M. (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.

Kalton, G. (1983). *Introduction to Survey Sampling.* Newbury Park, CA: Sage Publications.

Li, G., & Yao, B. (2018). Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches. *International Journal Of Design, Analysis And Tools For Intergrated Circuits And Systems*, 7(1), 63-67.

Liu, J., Wu, J., Liu, S., Li, M., Hu, K. & Li, K. (2021) Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *Plos One* 16(2): e0246306. https://doi.org/10.1371/journal.pone.0246306.

Peréz, V., Aybar, C. & Pavía, J.M. (2022). Dataset of the COVID-19 lockdown survey conducted by GIPEyOP in Spain. *Data in Brief,* 40, https://doi.org/10.1016/j.dib.2021.107700.

Rivers, D. (2007, August). Sampling for web surveys. In Joint Statistical Meetings (Vol. 4).

Rueda, M., Ferri-García, R., & Castro, L. (2020). The R package Non-ProbEst for estimation in non-probability surveys. The R Journal, 12(1), 406-418.