STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–14.
Published online in International Academic Press (www.IAPress.org)

Improving representativeness in big data analysis through weighted machine learning methods: A case study on the logistic regression model

Lamyae Benhlima*, Mohammed El Haj Tirari

Laboratory of Methods Applied in Statistics, Actuaries, Finance and Quantitative Economics, National Institute of Statistics and Applied Economics, Morocco

Abstract
In the presence of Big Data, it is essential to recognize that despite the abundance of data, these often do not faithfully represent the target populations. Therefore, analyzing these vast datasets does not guarantee representativeness, as they are collected without proper sampling design. Integrating survey weights and auxiliary information into machine learning algorithms constitutes a major challenge in making the samples more representative of the overall population. Moreover, only a few statistical learning software packages offer options to include these weights in their estimation process. In this paper, we introduce a novel weighted configuration of the logistic regression algorithm and employ a bootstrap method to compare its performance against non-weighted models. Our contributions demonstrate the importance and relevance of incorporating different weights for instances and provide a practical approach for analysts in settings where traditional statistical learning tools fall short. This work bridges a critical gap in statistical learning, ensuring that conclusions drawn from large datasets are robust and generalizable

Keywords Big Data, Machine Learning, Instances weighting, Logistic Regression, Sampling techniques

DOI: 10.19139/soic-2310-5070-2015

1. Introduction

Machine learning techniques include various methods used to extract information from databases for predictive or decision-making purposes. The advancement of learning theory can be attributed to the development of data storage resources and the advent of extensive data resulting from the digitalization of daily activities, capturing all digital footprints [1, 29]. Dealing with these vast amounts of data has led to the emergence of several open-source solutions to facilitate the processing and analysis of large-scale data [28]. However, handling large volumes of data introduces theoretical and mathematical challenges in terms of data analysis, as large volumes of data may not necessarily represent the target populations accurately. Similarly, many traditional statistical algorithms encounter issues of stability and robustness when applied at larger scales, as statistical tests tend to become significant in the presence of substantial data volumes [2, 3].
The convergence of machine learning and statistics is essential to capitalize on their respective strengths fully. Within the realm of machine learning, a key objective is to obtain reliable parameter estimates, enabling trained models to produce results applicable to the target population. Hence, the representativeness of initial samples assumes a critical role in ensuring the credibility of the generated outcomes. However, it is vital to acknowledge that the analysis of massive datasets does not automatically guarantee exhaustive results, as such data are frequently collected without a systematic sampling or survey design. This inherent characteristic introduces the risk of

*Correspondence to: Lamyae Benhlima (Email: l.benhlima@insea.ac.ma). Laboratory of Methods Applied in Statistics, Actuaries, Finance and Quantitative Economics, National Institute of Statistics and Applied Economics,Rabat, Morocco (10000).

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

encountering massive yet biased datasets, which may only represent specific subpopulations. To bridge this gap between machine learning and statistics, methodologies need to be implemented that ensure the representativeness of samples, thereby yielding more robust and generalizable conclusions for the target population. The integration of suitable weighting and sampling techniques serves as a pivotal component in this process, facilitating the incorporation of specific individual characteristics and ultimately enhancing the reliability of the findings.
As previously mentioned, ensuring the representativeness of the sample is essential for generalizing results to the target population. However, limited research has been conducted to assess the impact of ignoring sampling weights in statistical learning methods and its implications on the ability to draw conclusions applicable to the target populations[26, 27]. In the literature, two main approaches for incorporating weights in the analysis of machine learning models are discussed. The first approach involves post-hoc recalculation [4], where the model predictions are re-weighted after the initial model training to account for the weights. This approach allows the application of existing machine learning algorithms without significant modifications while considering the specific characteristics of the sample. On the other hand, the second approach involves directly incorporating weights into the process of estimating model parameters, thereby specifying the modifications to the inferential properties of these models [5, 6].
Our study falls within the scope of the second approach and aims to assess changes in the inferential properties of the logistic regression model by incorporating different weights for instances. Initially, we identified and implemented modifications to the inferential properties of the logistic regression model at various stages of configuration and evaluation through the incorporation of weight vectors. Subsequently, we conducted a series of simulations to train the learning model with and without weighting. Finally, we adopt a bootstrap-based approach to examine whether there exists a statistically significant difference between the outcomes of the weighted and unweighted models. This analysis allows us to determine the appropriateness of the weighting approach for a given problem.

2. Methodology

Traditional inferential tools for predictive models have been developed under the assumption of data obtained from a simple random sample, where each individual is given equal weight. This assumption implies that all individuals have equal importance in the inference process and represent the population to the same degree. However, in real-world scenarios, data is often collected through complex sampling designs, and individuals may have different probabilities of being selected for the sample. Introducing sampling weights into the core of the predictive model estimation process aims to address this complexity and provide a more accurate representation of the underlying population. In this context, sampling weights are used to differentiate the contributions of individuals in the sample. Individuals with lower weights have a diminished influence on the estimation and evaluation outcomes compared to those with higher weights. By incorporating distinct weightings for individuals within statistical learning algorithms, our objective is to ensure the representativeness of the sample and improve the precision of the produced estimators.

Specifically, we focus on the logistic regression model and examine its inferential properties after integrating sampling weights. Our proposed approach involves incorporating the weighting methodology into the logistic regression estimation process, effectively utilizing weighted likelihood. Each individual in the sample is assigned a specific weight value that encapsulates the characteristics of the sample design from which they were drawn. The allocation of these weights is determined by various aspects, such as the probability of inclusion, population size, sampling technique, stratification, and the availability of auxiliary information. By accounting for sampling weights, we can better account for the complexities of the data collection process and ensure that our statistical learning model is better equipped to make accurate and reliable predictions for the target population.

2.1. Logistic regression

Logistic regression is a widely used statistical method for modeling binary discrete dependent variables[30]. It involves one or more independent variables, which can be either continuous or categorical. The primary goal of

logistic regression is to estimate the probability of an event occurrence, yielding a predicted probability within the interval of 0 to 1. This probability can be utilized for binary classification or decision-making purposes. In this study, we measure a dichotomous response variable denoted as $Y_i, \quad i = 1, 2, \ldots, n$ for a group of n individuals, which can take two values, 0 and 1. Additionally, we have K explanatory variables (either qualitative or quantitative) denoted as $x_0, x_1, \ldots, x_K$. If an individual i belongs to the positive class, Y_i takes the value 1; otherwise, if the individual belongs to the negative class, Y_i takes the value 0. Our objective is to create Y as a function of the variables x_i using a logistic link while ensuring the fulfillment of specific assumptions [7]:

- N independent observations are denoted as $(Y_1, x_1), \ldots, (Y_n, x_n)$ where $x_i = (x_{i0}, \ldots, x_{iK})$
- Each Y_i given x_i follows a Bernoulli distribution with probability:

$$p_i = P(Y_i = 1 \mid x_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \tag{1}$$

with $\beta' x_i = \beta_0 x_{i0} + \ldots + \beta_K x_{iK}$, where $(\beta_1, \ldots, \beta_K)$ are unknown real parameters and $x_{i0} = 1$.

Thus, we aim to model the probability that Y takes the value 1 :

$$E(Y_i = 1/x_i) = p_i = P(Y_i = 1/x_i) \tag{2}$$

The logistic regression model, a specific type of generalized linear model, employs the logit transformation function logit $g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$, where p_i represents the probability of an event occurring. This transformation facilitates the establishment of a linear relationship between the dependent variable and the independent variables, as follows:

$$g(p_i) = \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_k \tag{3}$$

The coefficients $(\beta_1, \ldots, \beta_K)$ are estimated analytically using the likelihood method. A key objective of this research is to investigate the impact of incorporating sampling weights, which directly affect the formulation of the likelihood function. As a result, the introduction of sampling weights has significant implications for the equations used in maximum likelihood estimation.

Let $Y_i = (y_1, y_2, \ldots, y_n)$ be a simple random sample of a binary random variable. Y_i (given x_i) has the marginal density function over $\{0, 1\}$:

$$f(y_i/x_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{4}$$

Under the assumption of independence, the likelihood function is the product of the n marginal densities:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta' x_i}}\right)^{1-y_i} \tag{5}$$

The natural logarithm of the likelihood, known as the Log-Likelihood, is expressed as follows:

$$L(\beta) = \sum_{i=1}^{n} [y_i \beta' x_i - \log(1 + e^{\beta' x_i})] \tag{6}$$

By maximizing the log-likelihood function, we obtain regressions coefficients estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$. This method involves assigning $\hat{\beta}$ the value that maximizes the likelihood, or equivalently, the logarithm of the likelihood. However, this approach may lead to poorly performing classifiers due to the phenomenon of overfitting. In this particular case, it is common to use regularization techniques by adding penalty constraints to the log-likelihood function.

Two main approaches of regularization are widely used: the Ridge method proposed by Hoerl and Kennard [8], which adds a penalty term proportional to the sum of squares of the coefficients. This penalty, known as L2

regularization, helps control the model complexity by shrinking the coefficient values towards zero, preventing overfitting, and enhancing the model's ability to generalize to new data. Additionally, the Lasso method (Least Absolute Shrinkage and Selection Operator) proposed by Tibshirani [9], adds a penalty term proportional to the absolute values of the coefficients. This L1 regularization technique is effective in feature selection, as it encourages sparsity in the model by driving some coefficients to exactly zero, leading to a more interpretable and concise model. These regularization techniques play a crucial role in achieving better-performing and more stable models in various machine learning applications. The penalized likelihood is defined as follows:

$$L(\beta) = \sum_{i=1}^{n} [y_i \beta' x_i - \log(1 + \exp(\beta' x_i))] - \vartheta(\beta) \tag{7}$$

With,

$$\vartheta(\beta) = \begin{cases} \|\beta\|_1 = |\beta_0| + |\beta_1| + \ldots + |\beta_K|, \text{Lasso regularization} \\ \|\beta\|_2 = \sqrt{{\beta_0}^2 + {\beta_1}^2 + \ldots + {\beta_K}^2}, \text{Ridge regularization} \end{cases} \tag{8}$$

Here, $\lambda \geq 0$ is a regularization parameter controlling the amount of shrinkage. By incorporating the penalty term, we seek the value of β that nullifies the derivative of the penalized log-likelihood function (6). This leads to obtaining K equations, known as score equations, represented as follows:

$$\frac{\partial L(\beta)}{\partial \beta_k} = \begin{cases} \sum_{i=1}^{n} x_{ik} \left(y_i - \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)} \right) - \lambda \operatorname{sign}(\beta_k), \\ \qquad \text{if } \vartheta(\beta) = \|\beta\|_1 \\ \sum_{i=1}^{n} x_{ik} \left(y_i - \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)} \right) - 2\lambda \beta_k, \\ \qquad \text{if } \vartheta(\beta) = \|\beta\|_2 \end{cases} \tag{9}$$

The maximum likelihood estimator β' is obtained by solving a set of K nonlinear score equations. As a result, finding their solution necessitates the use of iterative numerical analysis methods. One commonly used technique for solving these equations is the Gradient Descent method [10]. This method is based on iterative optimization, aiming to determine the values of β that minimize the cost function associated with the logistic regression model.

2.2. Formulation of weighted likelihood

As mentioned earlier, weighted logistic regression involves adjusting the expression of the log-likelihood function to account for the weights assigned to each individual. This approach allows estimating model parameters by assigning weights to individuals based on their degree of representativeness or their importance. By incorporating these weights into the likelihood function, weighted likelihood enables obtaining parameter estimates that appropriately reflect the characteristics of the population. Weighted likelihood has proven to be a valuable technique in various statistical applications, especially when dealing with imbalanced and rare data [11]. The authors have proposed a new version of the weighted logistic regression algorithm (RE-WLR), demonstrating its robustness in handling large quantities of imbalanced data. Additionally, weighted likelihood is an effective approach for minimizing the mean squared error and correcting the bias of parameter estimates when the model is misspecified where the data follows a complex mixture of distributions [12]. The application of the weighted likelihood approach has also yielded promising results when dealing with outliers and their impact on estimation quality. In this context, researchers have introduced weighted score functions to replace the conventional ones. These new score functions incorporate weights that depend on the residuals, which helps mitigate the adverse effects of unusual observations. This adjustment contributes to more reliable and robust estimations, enhancing the overall effectiveness of the method in handling challenging data scenarios [22]. This research aims to demonstrate the pertinence of the weighted likelihood approach in the context of complex sampling, where the goal is to ensure sample representativeness. By incorporating survey weights assigned to each individual, this method ensures the production of robust estimates that effectively capture the diverse characteristics of the population.
Let $W = (w_1, w_2, \ldots, w_n)$ be the weight vector assigned to the n observed individuals. The weighted likelihood is

expressed as follows [13]:

$$L_W(\beta) = \prod_{i=1}^{n} [p_i^{y_i}(1-p_i)^{1-y_i}]^{w_i} \tag{10}$$

The weighted log-likelihood is obtained by taking the natural logarithm of equation (10):

$$L_W(\beta) = \sum_{i=1}^{n} w_i[y_i\beta' x_i - \log(1+\exp(\beta' x_i))] - \lambda\vartheta(\beta) \tag{11}$$

Thus, when individuals are weighted by the weight vector W, the estimator for β is obtained by maximizing the weighted likelihood:

$$\frac{\partial L_W(\beta)}{\partial \beta_k} = \begin{cases} \sum_{i=1}^{n} w_i \left(y_i - \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}\right) x_{ik} \\ -\lambda \operatorname{sign}(\beta_k), \qquad \text{if } \vartheta(\beta) = \|\beta\|_1 \\ \\ \sum_{i=1}^{n} w_i \left(y_i - \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}\right) x_{ik} \\ -2\lambda\beta_k, \qquad \text{if } \vartheta(\beta) = \|\beta\|_2 \end{cases} \tag{12}$$

Consequently, the crucial aspect of considering weighting in the logistic regression model involves the use of weighted likelihood and the resolving weighted score equations using numerical analysis techniques such as the Gradient Descent method.

3. Approach proposed

Statistical learning techniques focus on modeling variables to create predictive models, typically developing inferential tools under the assumption that the data originate from a simple random sample, where each individual in the sample is assigned equal weight. However, real-world data often come from complex sampling methods, where individuals may have differing levels of importance or relevance. To enhance the representativeness of samples and incorporate a priori knowledge, it is crucial to integrate weighting and calibration techniques that assign different weights to individuals. This adjustment is significant as it can substantially affect the inferential properties of the predictive models used in statistical learning. While some statistical learning software packages already support weighted methods, there remains a critical need for more extensive integration of these methods across a broader range of machine learning software tools. Such integration is essential for developing robust and equitable machine learning systems capable of effectively handling the complexities of real-world data across various domains.

3.1. Weight generating approach

Individual weighting has been extensively studied in scientific articles by d'Horwitz and Thompson [14]and by Hansen and Hurwitz [15], involving weighting units by the inverse of their inclusion probabilities. The weight assigned to an individual is determined by the sampling design and is interpreted as the number of the target population represented by the sampled individual, including themselves. Therefore, survey weights are expressed as $w_i = \frac{1}{\pi_i}$, where π_i represents the probability of selection for individual i.
In the context of machine learning datasets where sampling weights are often unknown, we employ a specific procedure to generate artificial weights for training samples. The approach we propose involves assuming that the test sample data is selected using simple random sampling, which is equivalent to not applying weights for parameter estimation. Subsequently, we refine these weights through calibration technique, ensuring that they align with the known totals of specific auxiliary variables [16],and incorporate them into the parameter estimation process of the machine-learning model. By employing the described approach to generate and refine sampling weights, we can effectively incorporate the appropriate weighting information. This step is crucial as it enhances the sample's representativeness and our models' accuracy, leading to better analysis and more reliable inference.
In the context of machine learning datasets where sampling weights are often unknown, we employ a specific

procedure to generate artificial weights for training samples. The approach we propose involves assuming that the test sample data is selected using simple random sampling, which is equivalent to not applying weights for parameter estimation. Subsequently, we refine these weights through calibration technique, ensuring that they align with the known totals of specific auxiliary variables [16], and incorporate them into the parameter estimation process of the machine-learning model. This approach is distinctive because, unlike traditional weighting methods, it uses known population information not included in the dataset, making the sample more representative. By employing this approach, we effectively incorporate appropriate weighting information, crucial for enhancing our models' accuracy and leading to better analysis and more reliable inference.

Especially in classification problems, it is crucial to ensure that the classes are well-represented in the sample. To address this concern, instead of a simple random sampling design, we have chosen a stratified sampling design, which involves selecting independent samples from each stratum based on their respective sizes and the overall population size.Within each stratum $(U_1, U_2, \ldots, U_H)$ characterized by sizes $(N_1, N_2, \ldots, N_H)$, we denote these independent samples as S_h with size n_h, where each n_h is determined by the characteristics of the specific stratum. The total population size N is defined as the sum of the sizes of all strata, given by $N = \sum_{h=1}^{H} N_h$.
The Neyman allocation method [17] is employed to determine the sample's sizes. This widely used approach in survey sampling theory optimizes the estimator of the total of the response variable at the population level by allocating sample sizes optimally across different strata. The objective is to minimize variance while maintaining a fixed overall sample size, denoted as n.

$$\begin{cases} \min\limits_{n_h} V_P\left[(\hat{t}_y\pi)\right] = \sum\limits_{h=1}^{H} \dfrac{N_h^2(1-f_h)}{n_h} S_{y,h}^2 \\ s/c : n = \sum\limits_{h=1}^{H} n_h \end{cases} \tag{13}$$

The resulting allocation from the minimization program is as follows:

$$n_h = n \cdot \frac{N_h \cdot S_h}{\sum_{h=1}^{H} N_h \cdot S_h} \tag{14}$$

Where,

$$S_{y,h}^2 = \begin{cases} \dfrac{1}{N_h - 1} \sum\limits_{k \in U_h} (y_k - \mu_{yh})^2 \\ \dfrac{N_h}{N_h - 1} P_h(1 - P_h) \end{cases} \tag{15}$$

The first equation represents the variance of the variable of interest when it is continuous, while the second one applies when it is binary.
With,

$$\mu_{yh} = \frac{1}{N_h} \sum_{k \in U_h} y_k \tag{16}$$

$$P_h = \frac{1}{N} \sum_{k \in U} y_{kh} \tag{17}$$

Thus, the weights obtained from the Optimal Stratified Sampling design are defined as $w_i = N_h/n_h$ for each individual i belonging to stratum h. Once the sampling weights have been determined, the next step involves applying calibration techniques to address potential estimation bias and improve alignment between the sample characteristics and those of the entire population. Calibration is a critical aspect of the estimation process, as it adjusts the weights based on auxiliary information available for the population. This method effectively corrects

any discrepancies between the sample and the target population, thereby enhancing the sample's representativeness by reducing biases introduced during the initial sampling design [18]. This adjustment ensures that the sample provides a more accurate representation of the population's characteristics, leading to improved precision in statistical estimates and strengthening the validity of inferences drawn from the sample. The calibration weight adjustment was performed following the Samplics approach [19].

3.2. Comparing weighted and unweighted models using the Bootstrap method

To evaluate the relevance of incorporating weights, we propose a bootstrap approach aimed at assessing whether there is a significant difference between the estimators generated by the model with and without weighting. The challenge stems from the unpredictability of the weights: they are random variables with generally unknown distributions. this distribution depends on the relationship between the weights and the other variables of the model. Therefore, a method that does not require prior knowledge of the weight distribution, namely the Bootstrap method, appears suitable for addressing this particular situation. The advantage of this method lies in its adaptability to parametric models, liberating us from the constraints of an unknown weight distribution.

Fundamental concept of the Bootstrap method : The Bootstrap method, initially introduced by Efron [20]. is a powerful technique that allows for the estimation of unknown characteristics of a population based on a set of samples drawn from it. This approach involves performing random sampling with replacement from a data sample to approximate the unknown distribution of a statistic. By analyzing the empirical distribution derived from multiple resampling iterations, we can estimate the distribution of the statistic of interest and draw inferences about the underlying population. Efron and Tibshyrani have conducted extensive research on the bootstrap, providing comprehensive explanations of this approach and demonstrating its application in regression models [21]. The Bootstrap concept revolves around the process of repeatedly drawing a large number of resamples from the original data to create an empirical bootstrap distribution of the statistic under consideration. This methodology offers a reliable approximation of the true distribution of the statistic, which is initially unknown. Multiple bootstrap variants are available, varying in how the bootstrap samples are generated and utilized. Among these, the most common approach is the empirical bootstrap, also known as pairwise bootstrap, where bootstrap samples are constructed by resampling the data without making any specific assumptions about the underlying distribution's form.

Bootstrap comparison of weighted and unweighted models: In this study, we aim to compare the results of weighted and unweighted models. Let $(Y_1, x_1) \dots (Y_n, x_n)$ represent a sample associated with a weight vector $W = (w_1, w_2, \dots, w_n)$. No assumptions are made about the weights, and the objective is to test the parameters of the logistic regression model under both scenarios. The estimators of β are denoted as $\hat{\beta}$ and $\hat{\beta}_w$ for the unweighted and weighted models, respectively. To compare the two estimators, we follow these steps:

- **Step 1: Subsampling**
  We create a new subsample, denoted as $(Y_1, x_1)^* \dots (Y_n, x_n)^*$, by performing random sampling with replacement from the original sample. With this new subsample, we apply the logistic regression model (both with and without weights) to obtain two new estimators of β denoted as $\hat{\beta}^*$ and $\hat{\beta}^*_w$.
- **Step 2: Estimator replication**
  We repeat the first step B times, resulting in B values of $\hat{\beta}^*$ and B values of $\hat{\beta}^*_w$.
- **Step 3: Comparison of results with and without weighting**
  We compare the means of the estimators $\hat{\beta}^*$ and $\hat{\beta}^*_w$ based on the respective B observations. We test the null hypothesis $H_0 : \overline{\hat{\beta}^*} = \overline{\hat{\beta}^*_w}$, which is asymptotically equivalent to testing $H_0 : E(\hat{\beta}) = E(\hat{\beta}_w)$. Rejecting the equality of means will lead us to conclude that there is a significant difference between the two models.

4. Simulation analysis

The simulation has been carried out utilizing datasets sourced from the UCI Machine Learning Repository [24, 25, 23, 31, 32]. The subsequent Table 1 presents an overview of these datasets, outlining their essential information:

Table 1. Datasets informations

Datasets	Instances	Associated Task
Census Income	48,842	Predict whether income exceeds $50,000/year
CDC Diabetes Health Indicators	253,680	Predict diabetes patients
Default of Credit Card	30,000	Predict credible clients
Mushromm	8124	Predict the edibility of a mushroom
Spambase	4601	Predict spam emails

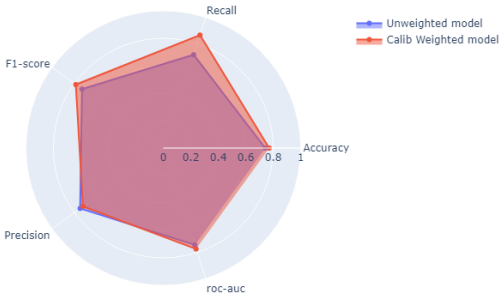
In the context of predictive modeling, data preparation and preprocessing include several crucial steps. The first step involves identifying and handling outliers, which are unusual values in the data that could significantly influence the results.Then, we proceeded with rebalancing the datasets using the undersampling technique to minimize bias in favor of the majority classes. Subsequently, two crucial processes, data encoding, and standardization are essential. Data encoding is used to convert categorical variables into appropriate numerical formats, while standardization is employed to scale numerical variables to a common range, facilitating easier comparison and interpretation of the data. Following these preparatory steps, the data is divided into training and testing sets. Enabling the evaluation of the model's performance on unseen data, to accurately assess its ability to generalize.

As the survey weights for each individual were unavailable in all datasets, we adopted an optimal stratified sampling approach to create a representative sample of the entire population. The main goal was to generate a weight vector to study the impact of considering the differential weighting of individuals on the learning model. To accomplish this, we implemented a stratified sampling design that ensured a balanced representation of each stratum in the sample. After drawing the sample, to improve the precision of the estimators, the calibration technique was employed to adjust the weight vector. This iterative process modifies the weights until the estimates from the calibrated sample align optimally with the true population values. To ensure consistency and reliability, we repeated this adjustment for different samples (100 samples in total) for each dataset. Through this rigorous approach, we could precisely evaluate the influence of survey weights on the learning model, leading to relevant and significant findings regarding the importance of this differentiated weighting approach.

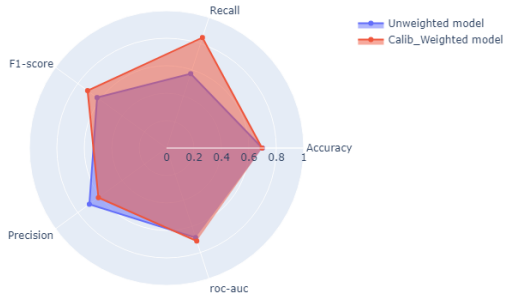
5. Results

To evaluate the logistic regression models' performance with and without weighting, we devised a new algorithm configuration that incorporates survey weights. After training both models, we evaluated their performance on the test data and compared the results. In the first scenario, where we implemented weighting, we calculated the performance indicators while considering the survey weights during the estimation process. The predicted results were calculated using probabilities generated by a model configured with weighting, effectively capturing the model's performance with a primary focus on the weights. This approach allows us to achieve results that have broader applicability. In contrast, in the second scenario, representing the unweighted approach, the observed results were derived from unweighted data, disregarding the individual weights. The predicted results also relied on probabilities generated by a model configured without weighting, reflecting the performance of a more basic model that is primarily aware of the sample itself. This scenario highlights an improper use of the data as it fails to account for variations in the distribution of variables between the study sample and the target populations. We visually represented the performance indicators for both scenarios in Figure 1:

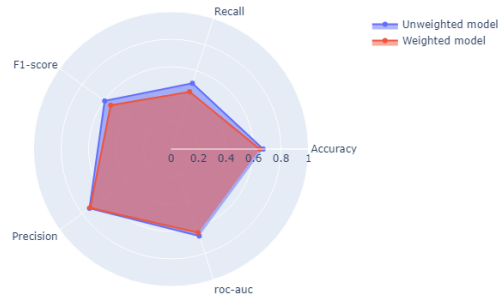
In our classification problems, we place special emphasis on accurately predicting individuals belonging to the positive class. To evaluate our model's effectiveness, we focus on two crucial performance indicators:



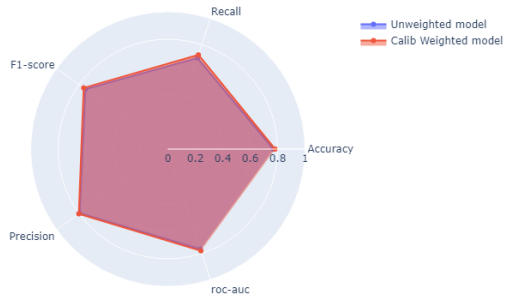
(a) Census Income dataset



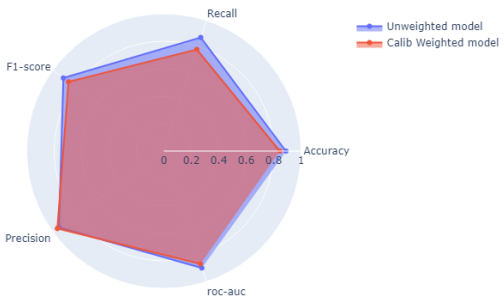
(b) Diabetes dataset



(c) Default of Credit Card



(d) Mushroom



(e) Spambase

Figure 1. Comparison of performance metrics of the logistic regression model with and without weighting

- **Recall**: This metric measures the model's capability to correctly identify true positives in relation to the total number of actual positive instances in the dataset.

- **F1-score**: This metric combines sensitivity and precision into a single score that considers the trade-off between detecting true positives and minimizing false positives.

Table 2. Evaluation of Differences in Performance Indicators between Weighted and Unweighted Models across 100 generated Samples

	Recall (%)				F1 score(%)			
DS[1]	M1[2]	M2[3]	Er1[4]	P1[5]	M3[2]	M4[3]	Er2[4]	P2[5]
CI	75.14	86.55	+11.41	2.38e-34	76.10	78.10	+2.00	4.89e-24
DHI	56.94	84.60	+27.7	1.18e-34	62.62	71.19	+8.75	1.20e-34
DCC	50.40	43.90	-6.5	4.9e-13	59.65	54.27	-5.38	1.86e-15
SPAM	87.03	77.87	-9.16	4.7e-27	90.56	85.91	-4.65	1.3e-22
Mushroom	69.76	72.16	+2.4	1.65e-3	74.25	75.87	+1.62	6.75e-4

[1] Datasets(CI:Census Income, DHI: Diabetes Health Indicaors,DCC: Default of Credit Card clients).
[2] M1 and M3 refer to the mean values of the Recall score and the F1 score , respectively, according to the unweighted model.
[3] M2 and M4 refer to the mean values of the F1 score and Recall score, respectively, according to the weighted model.
[4] Er1 and Er2 denote the difference between the mean values of the Recall score and F1 score, respectively, for the weighted model compared to the unweighted model.
[5] P1 and P2 represent the p-values from Mann-Whitney test comparing mean values of Recall and F1 scores, respectively, between the weighted and unweighted models.

As previously stated, the model training was conducted both with and without weighting across the 100 generated samples. Employing the non-parametric Mann-Whitney test, we established the presence of significant differences (p-values below the threshold of 0.05, as indicated in Table 2) between the means of the calculated indicators across various scenarios.These results robustly support our hypothesis that the introduction of sampling weights significantly influences the estimated parameters and, consequently, the performance indicators. This reveals how weighting effectively enhances model recall and the F1-score, thereby improving the model's ability to accurately identify true positives. These outcomes support our hypothesis that the introduction of sampling weights influences the estimated parameters and, consequently, the performance indicators.
To thoroughly ascertain statistically significant differences between the results of the two scenarios, we adopt the previously mentioned bootstrap method, following a systematic process. This begins with generating a sample through an optimal stratified sample design. Subsequently, employing the bootstrap technique with a resampling size of 50 creates multiple sub-samples from the original dataset. For each sub-sample, the weighted logistic regression model is applied to represent both scenarios and subsequent calculation of average estimated parameters is carried out. The comparison between the weighted and non-weighted models is accomplished through the non-parametric Mann-Whitney test, enabling the determination of any statistically significant differences. Consequently, informed decisions regarding the relevance of including survey weights within the study context can be made. For each dataset, we randomly select one of the generated samples to assess the difference between the weighted and non-weighted model using the bootstrap technique. The Mann-Whitney test reveals the rejection of the null hypothesis of equal means between the weighted and non-weighted models coefficients (see Appendix A). This highlights the importance of prioritizing the weighted model, which offers more accurate estimates of unknown parameters, thereby making it a suitable and reliable choice for representing the target population. These results underscore the significance of incorporating survey weights and emphasize their potential impact on improving the model's generalizability for real-world applications.

6. Discussion and conclusion

The meticulous pursuit of sample representativeness and the consideration of selection bias are imperative in ensuring the reliability of statistical learning models. Motivated by the challenge of achieving true representativeness in the era of Big Data, our study introduces a pivotal strategy: the adoption of sample weights. This strategy refines the accuracy of analyses, effectively bridging the gap between the sample and the true population. By incorporating this approach, we enhance the robustness and reliability of the results obtained, providing a powerful adjustment mechanism through initial weight assignment and subsequent calibration leveraging a priori knowledge. Incorporating sample weights into statistical learning is more than a procedural formality; it is a methodological necessity that allows analysts to capture the nuanced characteristics of individuals and achieve a more faithful representation of target populations. This strategic approach is crucial for mitigating biases associated with non-representative extensive datasets, ultimately bolstering the credibility of results derived from these techniques.

Our findings emphasize significant differences in logistic regression results, covering both model parameters and performance indicators, observed between the weighted and unweighted models. These differences highlight the crucial role of integrating sample weights and prior knowledge into prediction algorithms, ensuring the robust generalization of their results. While using a differential weighting system for individuals offers practical advantages, it necessitates careful consideration due to the uncontrolled variance introduced by weighted estimators, which could impact result precision. Practitioners must therefore make informed decisions based on how well model outcomes align under both weighted and unweighted scenarios. Additionally, selecting appropriate calibration variables for weighting is fraught with challenges, as they must be relevant to the target variables. These limitations render weighted models particularly sensitive to changes in data representation and quality, underscoring the need for cautious application in real-world scenarios.

In summary, the decision to use weighted or unweighted models hinges on the similarity of results obtained in both scenarios. If outcomes align closely, opting for the simplicity of an unweighted model is preferable to simplify the analysis and enable exact inference through tests. However, when notable differences arise, it is strongly advisable to use sample weights to achieve unbiased or lower-bias estimators. As we consider the future directions for this research, the potential extensions of our weighted approach to other machine learning algorithms appear particularly promising. While this study focused primarily on logistic regression, future work could explore the implementation of similar weighting techniques in algorithms such as decision trees, support vector machines, and neural networks. Each of these models may benefit differently from the introduction of weights, particularly in their ability to handle various data imbalances and biases. Furthermore, investigating the applicability of our approach across different domains such as healthcare, finance, and social media analytics and with varying data characteristics could provide deeper insights into its versatility and effectiveness. Such explorations would not only validate the robustness of our methodology but also enhance its relevance for practical applications in fields where data representativeness is critical.

Appendix A Mann-Whitney test results

Table 3. Mann-Whitney test results for Census Income dataset

Model Coefficient	P-Value	Test Decision
β_0	0.00	Rejecting null hypothesis of equal means
β_1	0.00	Rejecting null hypothesis of equal means
β_2	0.00	Rejecting null hypothesis of equal means
β_3	0.00	Rejecting null hypothesis of equal means
β_4	0.00	Rejecting null hypothesis of equal means
β_5	0.00	Rejecting null hypothesis of equal means
β_6	0.00	Rejecting null hypothesis of equal means
β_7	0.00	Rejecting null hypothesis of equal means
β_8	0.00	Rejecting null hypothesis of equal means
β_9	0.00	Rejecting null hypothesis of equal means
β_{10}	0.00	Rejecting null hypothesis of equal means
β_{11}	0.00	Rejecting null hypothesis of equal means
β_{12}	0.00	Rejecting null hypothesis of equal means
β_{13}	0.00	Rejecting null hypothesis of equal means
β_{14}	0.00	Rejecting null hypothesis of equal means
β_{15}	0.00	Rejecting null hypothesis of equal means
β_{16}	0.00	Rejecting null hypothesis of equal means
β_{17}	0.00	Rejecting null hypothesis of equal means
β_{18}	0.00	Rejecting null hypothesis of equal means
β_{19}	0.00	Rejecting null hypothesis of equal means
β_{20}	0.00	Rejecting null hypothesis of equal means
β_{21}	0.00	Rejecting null hypothesis of equal means
β_{22}	0.00	Rejecting null hypothesis of equal means
β_{23}	0.00	Rejecting null hypothesis of equal means
β_{24}	0.00	Rejecting null hypothesis of equal means
β_{25}	0.00	Rejecting null hypothesis of equal means
β_{26}	0.00	Rejecting null hypothesis of equal means
β_{27}	0.00	Rejecting null hypothesis of equal means

Table 4. Mann-Whitney test results for CDC Diabetes Health Indicators dataset

Model Coefficient	P-Value	Test Decision
β_0	0.00	Rejecting null hypothesis of equal means
β_1	0.00	Rejecting null hypothesis of equal means
β_2	0.00	Rejecting null hypothesis of equal means
β_3	0.00	Rejecting null hypothesis of equal means
β_4	0.00	Rejecting null hypothesis of equal means
β_5	0.00	Rejecting null hypothesis of equal means
β_6	0.00	Rejecting null hypothesis of equal means
β_7	0.00	Rejecting null hypothesis of equal means
β_8	0.00	Rejecting null hypothesis of equal means
β_9	0.00	Rejecting null hypothesis of equal means
β_{10}	0.00	Rejecting null hypothesis of equal means

Table 5. Mann-Whitney test results for Default of Credit Card dataset

Model Coefficient	**P-Value**	**Test Decision**
β_0	0.00	Rejecting null hypothesis of equal means
β_1	0.00	Rejecting null hypothesis of equal means
β_2	0.00	Rejecting null hypothesis of equal means
β_3	0.00	Rejecting null hypothesis of equal means
β_4	0.00	Rejecting null hypothesis of equal means
β_5	0.00	Rejecting null hypothesis of equal means
β_6	0.00	Rejecting null hypothesis of equal means
β_7	0.00	Rejecting null hypothesis of equal means
β_8	0.00	Rejecting null hypothesis of equal means
β_9	0.00	Rejecting null hypothesis of equal means
β_{10}	0.00	Rejecting null hypothesis of equal means
β_{11}	0.00	Rejecting null hypothesis of equal means
β_{12}	0.00	Rejecting null hypothesis of equal means
β_{13}	0.00	Rejecting null hypothesis of equal means
β_{14}	0.00	Rejecting null hypothesis of equal means
β_{15}	0.00	Rejecting null hypothesis of equal means
β_{16}	0.00	Rejecting null hypothesis of equal means
β_{17}	0.04	Rejecting null hypothesis of equal means
β_{18}	0.00	Rejecting null hypothesis of equal means
β_{19}	0.00	Rejecting null hypothesis of equal means
β_{20}	0.00	Rejecting null hypothesis of equal means
β_{21}	0.00	Rejecting null hypothesis of equal means
β_{22}	0.00	Rejecting null hypothesis of equal means
β_{23}	0.00	Rejecting null hypothesis of equal means
β_{24}	0.00	Rejecting null hypothesis of equal means

Table 6. Mann-Whitney test results for Mushroom dataset

Model Coefficient	**P-Value**	**Test Decision**
β_0	0.00	Rejecting null hypothesis of equal means
β_1	0.00	Rejecting null hypothesis of equal means
β_2	0.00	Rejecting null hypothesis of equal means
β_3	0.00	Rejecting null hypothesis of equal means
β_4	0.00	Rejecting null hypothesis of equal means
β_5	0.00	Rejecting null hypothesis of equal means

Table 7. Mann-Whitney test results for Spambase dataset

Model Coefficient	P-Value	Test Decision
β_0	0.00	Rejecting null hypothesis of equal means
β_1	0.00	Rejecting null hypothesis of equal means
β_2	0.00	Rejecting null hypothesis of equal means
β_3	0.00	Rejecting null hypothesis of equal means
β_4	0.00	Rejecting null hypothesis of equal means
β_5	0.00	Rejecting null hypothesis of equal means
β_6	0.00	Rejecting null hypothesis of equal means
β_7	0.00	Rejecting null hypothesis of equal means
β_8	0.00	Rejecting null hypothesis of equal means
β_9	0.00	Rejecting null hypothesis of equal means
β_{10}	0.00	Rejecting null hypothesis of equal means
β_{11}	0.00	Rejecting null hypothesis of equal means
β_{12}	0.00	Rejecting null hypothesis of equal means
β_{13}	0.00	Rejecting null hypothesis of equal means
β_{14}	0.00	Rejecting null hypothesis of equal means
β_{15}	0.00	Rejecting null hypothesis of equal means
β_{16}	0.00	Rejecting null hypothesis of equal means
β_{17}	0.00	Rejecting null hypothesis of equal means
β_{18}	0.00	Rejecting null hypothesis of equal means
β_{19}	0.00	Rejecting null hypothesis of equal means
β_{20}	0.00	Rejecting null hypothesis of equal means
β_{21}	0.00	Rejecting null hypothesis of equal means
β_{22}	0.00	Rejecting null hypothesis of equal means
β_{23}	0.00	Rejecting null hypothesis of equal means
β_{24}	0.00	Rejecting null hypothesis of equal means
β_{25}	0.00	Rejecting null hypothesis of equal means
β_{26}	0.00	Rejecting null hypothesis of equal means
β_{27}	0.00	Rejecting null hypothesis of equal means
β_{28}	0.00	Rejecting null hypothesis of equal means
β_{29}	0.00	Rejecting null hypothesis of equal means
β_{30}	0.00	Rejecting null hypothesis of equal means
β_{31}	0.00	Rejecting null hypothesis of equal means
β_{32}	0.00	Rejecting null hypothesis of equal means
β_{33}	0.00	Rejecting null hypothesis of equal means
β_{34}	0.00	Rejecting null hypothesis of equal means
β_{35}	0.00	Rejecting null hypothesis of equal means
β_{36}	0.00	Rejecting null hypothesis of equal means
β_{37}	0.00	Rejecting null hypothesis of equal means
β_{38}	0.00	Rejecting null hypothesis of equal means

REFERENCES

1. Alpaydin, Ethem, *Introduction to machine learning*, The MIT Press, Cambridge, Massachusetts, 2020.
2. Carmichael, Iain and Marron, J. S., *Data science vs. statistics: two cultures?*, Japanese Journal of Statistics and Data Science, vol. 1, pp. 117–138, 2018, DOI: 10.1007/s42081-018-0009-3.
3. Breiman, Leo, *Statistical Modeling: The Two Cultures*, Statistical Science, vol. 16, pp. 199–215, 2001.
4. MacNell, Nathaniel and Feinstein, Lydia and Wilkerson, Jesse and Salo, Paivi M. and Molsberry, Samantha A. and Fessler, Michael B. and Thorne, Peter S. and Motsinger-Reif, Alison A. and Zeldin, Darryl C., *Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting*, PLOS ONE, vol. 18, pp. e0280387, 2023, DOI: 10.1371/journal.pone.0280387.
5. Mahdi Hashemi and Hassan A. Karimi, *Weighted Machine Learning*, Statistics Optimization and Information Computing, vol. 6, pp. 497–525, 2018, DOI: 10.19139/soic.v6i4.479.
6. Byrd, Jonathon and Lipton, Zachary C., *What is the Effect of Importance Weighting in Deep Learning?*, Paper presented at the 36th International Conference on Machine Learning, 2019, DOI: 10.48550/arXiv.1812.03372.
7. David W. Hosmer, Jr., *Applied Logistic Regression*, Wiley, New York, 2000.
8. Hoerl, Arthur E. and Kennard, Robert W., *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, vol. 12, pp. 55–67, 1970, DOI: 10.1080/00401706.1970.10488634.
9. Tibshirani, Robert, *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, pp. 267–288, 1996, DOI: 10.1111/j.2517-6161.1996.tb02080.x.
10. Zhang, Tong, *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, Paper presented at the Twenty-first International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, DOI: 10.1145/1015330.1015332.
11. Maalouf, Maher and Siddiqi, Mohammad, *Weighted logistic regression for large-scale imbalanced and rare events data*, Knowledge-Based Systems, vol. 59, pp. 142–148, 2014, DOI: 10.1016/j.knosys.2014.01.012.
12. Markatou, Marianthi, *Mixture Models, Robustness, and the Weighted Likelihood Methodology*, Biometrics, vol. 56, pp. 483–486, 2000, DOI: 10.1111/j.0006-341X.2000.00483.x.
13. Markatou, Marianthi and Basu, Ayanedranath and Lindsay, Bruce, *Weighted likelihood estimating equations: The discrete case with applications to logistic regression*, Journal of Statistical Planning and Inference, vol. 57, pp. 215–232, 1997, DOI: 10.1016/j.knosys.2014.01.012.
14. D. G. Horvitz and D. J. Thompson, *A Generalization of Sampling Without Replacement From a Finite Universe*, Journal of the American Statistical Association, vol. 47, pp. 663–685, 1952, DOI: 10.2307/2280784.
15. Hansen, Morris H. and Hurwitz, William N., *The Problem of Non-Response in Sample Surveys*, Journal of the American Statistical Association, vol. 41, pp. 517–529, 2004, DOI: 10.1080/01621459.1946.10501894.
16. Deville, Jean-Claude and Särndal, Carl-Erik, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, vol. 87, pp. 376–382, 1992, DOI: 10.1080/01621459.1992.10475217.
17. Tillé, Yves and Favre, Anne-Catherine, *Optimal allocation in balanced sampling*, Statistics & Probability Letters, vol. 74, pp. 31–37, 2005, DOI: 10.1016/j.spl.2005.04.027.
18. Valliant, Richard and Dever, Jill A. and Kreuter, Frauke, *Calibration and Other Uses of Auxiliary Data in Weighting*, in M. Broy and E. Denert (Eds.), *Practical Tools for Design*
19. Diallo, Mamadou, *samplics: a Python Package for selecting, weighting and analyzing data from complex sampling designs*, *Journal of Open Source Software*, vol. 6, pp. 3376, 2021, DOI: 10.21105/joss.03376
20. Efron, B., *Bootstrap Methods: Another Look at the Jackknife*, in S. Kotz and N.L. Johnson (Eds.), *Breakthroughs in Statistics*, Springer, New York, pp. 569–593, 1992, DOI: 10.1007/978-1-4612-4380-9-41
21. Bradley Efron and Robert J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993, DOI: 10.1201/9780429246593, pp. 105-121
22. Feifang Hu and James V. Zidek, *The Weighted Likelihood*, *The Canadian Journal of Statistics*, vol. 30, pp. 347–371, 2002, DOI: 10.2307/3316141
23. Yeh, I-Cheng, *default of credit card clients*, UCI Machine Learning Repository, 2016, DOI: 10.24432/C55S3H
24. Becker, Barry and Kohavi, Ronny, *Adult*, UCI Machine Learning Repository, 1996, DOI: 10.24432/C5XW20
25. Centers for Disease Control and Prevention, *CDC Diabetes Health Indicators*, UCI Machine Learning Repository, 2015, DOI: 10.24432/C53919
26. Clemmensen, LH and Kjærsgaard, RD, *Data representativity for machine learning and ai systems*, arXiv preprint arXiv:2203.04706, 2023.
27. Asudeh, Abolfazl and Jin, Zhongjun and Jagadish, HV, *Assessing and remedying coverage for a given dataset*, 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 554–565, 2019, IEEE.
28. Shahbazi, Nima and Lin, Yin and Asudeh, Abolfazl and Jagadish, HV, *Representation bias in data: a survey on identification and resolution techniques*, ACM Computing Surveys, vol. 55, no. 13s, pp. 1–39, 2023, ACM New York, NY.
29. Goswami, Tilottama and Sinha, GR, *Statistical Modeling in Machine Learning: Concepts and Applications*, Academic Press, 2022.
30. Goswami, Tilottama and Sinha, GR, *Statistical Modeling in Machine Learning: Concepts and Applications*, Academic Press, 2022.
31. UCI Machine Learning Repository, *Mushroom*, DOI:10.24432/C5959T, 1987.
32. Hopkins, Mark, Reeber, Erik, Forman, George, and Suermondt, Jaap, *Spambase*, UCI Machine Learning Repository, DOI:10.24432/C53G6X, 1999.