Working Paper Series

Predicting Poverty with Missing Incomes

Paolo Verme

ECINEQ 2023 642

ECINEQ 2023 642
2023 March
www.ecineq.org

Predicting Poverty with Missing Incomes

Paolo Verme
World Bank

Abstract

Poverty prediction models are used by economists to address missing data issues in a variety of contexts such as poverty profiling, targeting with proxy-means tests, cross-survey imputations such as poverty mapping, or vulnerability analyses. Based on the models used by this literature, this paper conducts an experiment by artificially corrupting data with different patterns and shares of missing incomes. It then compares the capacity of classic econometric and machine learning models to predict poverty under these different scenarios. It finds that the quality of predictions and the choice of the optimal prediction model are dependent on the distribution of observed and unobserved incomes, the poverty line, the choice of objective function and policy preferences, and various other modeling choices. Logistic and random forest models are found to be more robust than other models to variations in these features, but no model invariably outperforms all others. The paper concludes with some reflections on the use of these models for predicting poverty.

Keyword: Income modeling, Income Distributions, Poverty Predictions

JEL Cassification: D31, D63, E64, O15

Predicting Poverty with Missing Incomes

Paolo Verme*

March 28, 2023

Abstract

Poverty prediction models are used by economists to address missing data issues in a variety of contexts such as poverty profiling, targeting with proxy-means tests, cross-survey imputations such as poverty mapping, or vulnerability analyses. Based on the models used by this literature, this paper conducts an experiment by artificially corrupting data with different patterns and shares of missing incomes. It then compares the capacity of classic econometric and machine learning models to predict poverty under these different scenarios. It finds that the quality of predictions and the choice of the optimal prediction model are dependent on the distribution of observed and unobserved incomes, the poverty line, the choice of objective function and policy preferences, and various other modeling choices. Logistic and random forest models are found to be more robust than other models to variations in these features, but no model invariably outperforms all others. The paper concludes with some reflections on the use of these models for predicting poverty.

Keywords: income modeling; Income Distributions; Poverty Predictions; Imputations.
JEL Codes: D31; D63; E64; O15.

*World Bank. The author is grateful to Bo Pieter Johannes Andree, Olivier Dupriez, Aivin Vicquierra Solatorio, Lidia Ceriani and David Newhouse for excellent comments on the first draft.

1 Introduction

The poverty rate, defined as the share of poor people in a given population, is an important indicator of well-being. It is one of the main indicators adopted by the UN Sustainable Development Goals (SDGs), it is used by International Financial Institutions (IFIs) for the global count of the poor, to classify countries according to their level of well-being, and allocate global financial resources. Estimates of poverty at the household level are also used by national and local governments to target populations in need of assistance and are a core instrument of social protection policies. An accurate estimate of poverty at the population or household level is a precondition for effective global, national and local welfare policies.

Accurate poverty measurement is not a simple exercise. It is based on sample surveys that collect information on monetary indicators such as income, consumption or expenditure. It is therefore a sample based *estimate* of the population poverty rate. These estimates suffer from a variety of measurement errors including sampling errors, misreporting on the part of respondents or interviewers, and unit or item non-response. Once the data are collected, statistical agencies may also apply alterations to the data that can potentially impair proper statistical estimates such as top coding, or deletion of outliers. No survey is exempted from at least some of these issues and some of these issues can affect a very large share of the sample. The poverty rate is measured almost invariably with money metrics that contain some degree of missing or unreliable items.

Several strands of the poverty measurement literature can be regarded as poverty prediction exercises designed to address missing data issues. Targeting exercises using proxy-means testing are designed to predict poverty when household income is not available (Coady et al., 2004, Brown et al., 2018, Glewwe, 1991, Baker and Grosh, 1994). Poverty profiles use single or multiple imputations to replace missing incomes for item non-response and use predicted and observed incomes to estimate poverty. Poverty mapping and cross-survey imputations exercises use prediction models to estimate poverty in surveys that do not have information on incomes (Elbers et al., 2003, Tarozzi and Deaton, 2009). Top or bottom incomes studies have used parametric and non-parametric methods to replace missing observations on the tails of a distribution to estimate poverty or inequality (Cowell and Victoria-Feser, 1996a, Korinek et al., 2007, Atkinson et al., 2011, Jenkins, 2017). Vulnerability assessments use poverty prediction models to estimate the probability of poverty for a hypothetical future period not yet observed (Morduch, 1994; Calvo and Dercon, 2013; Verme et al., 2016). More recently, poverty specialists have started to use machine learning methods to predict poverty when incomes are missing from parts

of the distribution (Hlasny et al., 2021) or by using these methods with innovative data (Blumenstock et al., 2015, Abelson et al., 2014, Jean et al., 2016, Mcbride and Nichols, 2018, Andree, 2021, Aiken et al., 2022, Aiken et al., 2023).

What all these models have in common is that they try to estimate poverty (or some measure of well-being) with a prediction model and in the presence of missing observations. The initial outcome variable of the prediction model may be income or poverty depending on the type of model. But the statistics of interest to estimate is poverty at the population or household level. We will generally refer to these models as poverty prediction models since these models either predict poverty or estimate poverty based on predicted incomes.

The key question with all these models is how to predict poverty accurately when some or all incomes are missing, a question that is hampered by at least two factors: 1) Incomes in surveys used for poverty measurement have been shown to be Missing Not At Random (MNAR) (Lillard et al., 1986, D'Alessio and Faiella, 2002, D'Alessio and Neri, 2015, Bollinger et al., 2019, Hlasny and Verme, 2021); 2) There is no real counterfactual. The true poverty rate is not observed because of the presence of missing observations.

Building on the literature cited above, this paper compares the performance of classic econometric[1] and machine learning models in predicting poverty with different missing observations shares and patterns and against the true poverty rate. This is done by generating a sample with no missing observations and corrupting this sample with various shares and patterns of missing observations. The performance of poverty prediction models can then be assessed with complete information on the full distribution of incomes, the true poverty rate and the specific missing data patterns. The paper also provides a framework to compare classic econometric and machine learning models.

The objective of the paper is to show how different classic econometric and machine learning models behave for predicting poverty when data, objective function (loss function), poverty lines, or various parameters and prediction strategies change. We are not striving to find the ultimate prediction model for the data at hand but understand how different prediction models respond to changes in these features using experimental data. The analysis is based on income data and the prediction models considered require income predictors to be observed for households that do not report incomes (item non-response). Results of this paper are not necessarily valid for other money metrics of well-being such as consumption or expenditure, or for other types of missing data issues such as unit non-response or top-coding.

[1]By "classic econometric" we mean standard OLS and Maximum Likelihood models such as logit or probit models". Formal definitions are provided further in the paper.

Results show that the quality of poverty predictions and the choice of the optimal prediction model are dependent on the distribution of observed and unobserved incomes, the poverty line, the choice of objective function and policy preferences, the choice of models' parameters, and the use or non use of various optimization strategies. Logistic and Random Forest models are more robust than other models to variations in these features but no model invariably outperforms all others. Logistic and Random Forest models seem to have an edge on other models because are better suited to predict incomes in the tails of distributions irrespective of data and objective functions. However, all models have specific adjustments that can help to improve predictions significantly.

The paper is organized as follows. The next section describes common missing data problems. Section 3 outlines how welfare economists have addressed this problem. Section 4 provides a consistent framework that can be used to compare classic econometric and machine learning models. Section 5 describes the data. Section 6 conducts an experiment with dummy data to compare the capacity of these models to predict the poverty rate accurately. Section 7 provides a series of robustness tests by varying data, parameters and preferences. Section 8 provides additional tests by calibrating the models. Section 9 concludes by summarizing the main findings and providing some initial indications on how these models can be used effectively.

2 The distribution of missing data

Estimating survey based statistics with missing data is a known challenge and addressing a missing data issue requires an understanding the nature of missing data. Statisticians (Rubin, 1976, Rubin and Little, 2020) distinguish between data Missing Completely At Random (MCAR), when there is no apparent law that regulates missing data; Missing At Random (MAR) where missing data of the outcome variable of interest are correlated with covariates of this outcome but not with the outcome itself (for example, when only men are not responding to income questions because of their gender, not their income), and Missing Not At Random (MNAR) where missing data of the outcome variable of interest are correlated with the outcome variable itself (for example, higher income households are less likely to respond to income questions in surveys because they have higher incomes).

More formally and following Rubin and Little (2020), let Y_{ij} be the complete data matrix and M_{ij} the indicator matrix representing missing observations where i represent observations and j represent variables. Then, the distribution of m_i conditional on y_i is $f_{M|Y}(m_i|y_i, \phi)$ with ϕ being the unknown parameters of the function that relates m to y. If M_{ij} does not depend on Y_{ij}, it is said that missing

data are Missing Completely at Random (MCAR). Let now $y(0)_i$ be the components of y_i that are observed for unit i, and $y(1)_i$ the components of y_i that are missing for unit i. A less restrictive assumption than MCAR is that m_i depends on y_i only through the observed components $y(0)_i$. This case is defined as observations Missing At Random (MAR). Finally, the missing data pattern is called Missing Not At Random (MNAR) if the distribution of m_i depends on y_i, which is only partially observed.

In the case of poverty measurement, the problem of non-randomness is particularly acute when poverty is measured with incomes, a standard practice in high and many middle-income countries. Missing incomes in surveys are known to be correlated with income itself in rich and poor countries alike (Atkinson et al., 2011, Hlasny and Verme, 2018a), and there is evidence from multiple countries that income non-responses are U-shaped with both lower and upper income households less likely to respond to questions in surveys (Lillard et al., 1986; Bollinger et al., 2019; D'Alessio and Faiella, 2002; D'Alessio and Neri, 2015). This fact makes income data MNAR and possibly also MAR, since some predictors of incomes are also likely to be associated with missing data. From a statistical standpoint, this is the most complex scenario for proper statistical estimates because the function that relates m_i to y_i (the ϕ parameters) is unknown. It is a scenario where the basic assumptions needed to use popular imputation methods such as multiple imputations are not met.

In our knowledge, the only strand of the poverty literature that has focused on this problem is a string of papers that uses a GMM method to estimate the probability of non-response in sample. This method estimates the function that relates m_i to y_i (the ϕ parameters) addressing the fundamental missing data problem. Authors who used this method found income non-responses to be strongly associated with income (Korinek et al. (2006); Korinek et al. (2007); Hlasny and Verme, 2018a; Hlasny and Verme, 2018b; Hlasny and Verme, 2021). When one has a large share of missing incomes and no means to test their pattern, one should assume that missing incomes are MNAR and that estimating the poverty rate on observed values only is likely to bias this estimation significantly.

While the problem of missing incomes is generally discussed in the context of in-survey imputations, it is equally relevant for targeting, cross-survey imputations, and vulnerability analyses for two reasons. One is that the original sample used for modeling incomes is also likely to suffer from missing incomes as most surveys do, which has implications for the predicted values out of sample. And second, even if all incomes are observed in the modeling sample, predicting out of sample amounts to predicting missing incomes for the totality of households (individuals) in the imputation sample. It is as if the two samples were compiled into one and the

incomes related to the imputation sample were missing. This can also be regarded as a missing data problem.

3 How poverty economists address missing data issues to predict poverty

There are at least five strands of the poverty measurement literature that treat the question of missing incomes with prediction models based on classic econometric methods: The literature on targeting and proxy-means testing, the literature on cross-survey imputations such as poverty mapping, the literature on top incomes, the literature on vulnerability to poverty, and the in-survey imputation literature. More recently, poverty economists have experimented with machine learning methods. We briefly review these strands of the literature in this order.

The literature on proxy-means testing used for targeting makes extensive use of prediction methods to estimate poverty for households where income is non-observed (Coady et al., 2004, Brown et al., 2018, Glewwe, 1991, Baker and Grosh, 1994). The idea is that one can predict poverty using a restricted set of observed socio-economic characteristics avoiding in this way extensive and expensive surveys on income or consumption. In this case, the prediction model is built on an existing survey representative of the population of interest. A short survey is then administered to potential beneficiaries to collect data on key income predictors as identified by the model. This information is, in turn, used to predict poverty for individual households or assign a score that can rank households according to their level of well-being. This literature has used standard OLS or Logistic models for the prediction model. It implicitly assumes that missing data from the sample used for the prediction model are not problematic and that targeting beneficiaries are extracted randomly from the same population covered by the prediction model. A proper discussion of missing data in works on proxy-means testing is rarely seen but the models used are poverty prediction models covering households with missing incomes.

Cross-survey imputation methods have been developed to estimate poverty when income or consumption data are missing from the survey of interest but can potentially be estimated using other surveys representative of the same population and including incomes. One example is small areas estimations also referred to as "poverty mapping". The idea behind poverty mapping is to use poverty predictors extracted from censuses to predict poverty at the micro geographical level using the coefficients of a prediction model estimated with survey data that contain incomes at the the macro level (Elbers et al., 2003, Tarozzi and Deaton, 2009). Although

this literature has developed rather independently of the multiple imputation literature in statistics[2], it uses multiple imputation methods and builds on these methods to specifically address the question of reduced variance among predicted values in the context of continuous dependent variable models. Similar imputation methods for the purpose of predicting poverty have also been used across years (Dang et al., 2019), different types of surveys such as consumption and labor force surveys (Douidich et al., 2016), or different types of data such as administrative and survey data (Dang and Verme, 2022). This literature has used continuous (Elbers et al., 2003) and categorical (Tarozzi and Deaton, 2009) dependent variable models to predict poverty. As for the proxy-means tests literature, this literature rarely discusses missing data patterns but addresses the important question of the correct estimation of the variance of predicted values.

The literature on top incomes has focused on the fact that top incomes are under represented in surveys and that a correct estimation of inequality in any given country needs to address this issue (Atkinson et al., 2011, Jenkins, 2017, Hlasny and Verme, 2021). This literature recognizes that missing observations are an increasing function of income and are, therefore, MNAR.[3] Several methods have been proposed to address this issue ranging from replacing top incomes with observations extracted from theoretical distribution functions such as Pareto (Cowell and Victoria-Feser, 1996b, Jenkins, 2017), to replacing top incomes with data external to the survey such as tax data (Atkinson et al., 2011), to reweighting observations using the inverse of the probability of non-response estimated from observed data (Korinek et al., 2007, Korinek et al., 2006, Hlasny and Verme, 2018a). Replacing observations with theoretical distributions or external data can be effective when missing observations are almost exclusively on the tails of a distribution but these methods are less efficient when missing observations are located closer to central values. Reweighting methods are more indicated to estimate missing observations all along the distribution and they also have the distinct advantage of estimating the probability of non-response, which is the function that relates m_i to y_i. However, in order to implement this method, one has to have non-response rates at a very disaggregated level, an information that is not always available to researchers. Unlike the proxy-means testing and cross-survey imputation literature, this literature focuses on the missing data question and inequality. However, these same methods have also been extended to the study of bottom incomes and poverty (Cowell and Victoria-Feser, 1996a; Hlasny

[2]Elbers et al. (2003) does not refer to the Rubin or Imbens literature while Tarozzi and Deaton (2009) refers to several of Rubin's papers but not to those that specifically addressed the cross-survey imputation question.

[3]Interestingly, this literature rarely refers to MNAR data explicitly.

et al., 2021).

Scholars working on vulnerability to poverty have also used prediction methods to gauge the probability of poverty in the future by simply estimating this probability with a OLS or Logit prediction model (Morduch, 1994; Calvo and Dercon, 2013; Verme et al., 2016). This literature has not been particularly concerned with either missing items as the top incomes literature or the error term as for the cross-survey imputation literature. However, it is similar to the case where the cross-survey imputation methods are applied to surveys administered in different years, with the important difference that predictions are made in sample and not out of sample.

All these strands of the literature may also use in-survey single or multiple imputations to estimate incomes for item non-response in the data used for modeling. This is where households are captured in the sample but do not reply to the income question. In this case, one can estimate incomes based on the other socio-economic characteristics observed with single or multiple imputation methods. This is also a standard practice used by practitioners working on poverty profiles.

More recently, machine learning methods have also been used by economists to predict poverty with a variety of innovative data such as mobile phone (Blumenstock et al., 2015), satellite imagery and remote sensing data (Abelson et al., 2014, Jean et al., 2016), or for targeting the poor (Mcbride and Nichols, 2018; Aiken et al., 2023). A global competition launched by the World Bank to predict poverty with machine learning algorithms provided some initial evidence on how these methods can help to improve on classic poverty prediction methods.[4] All these studies largely relied on standard ML methods including tree based methods, regularization, and neural networks or deep learning methods. These are the ML methods considered by this paper.

4 Baseline Framework for Comparing models

4.1 Three steps' predictions

As shown in the previous section, one important distinction that the different strands of the poverty prediction literature share is the distinction between continuous and discrete (dichotomous) dependent variable models. These two types of models are applicable in the context of classic econometric and machine learning models lending themselves to be a useful framework to compare these different approaches to poverty predictions. This section clarifies the steps required to classify households into poor

[4]See details of this competition on https://www.drivendata.co/blog/poverty-winners/.

and non-poor households and the difference between these two sets of models.

To illustrate these differences, we use a simple OLS model based on a continuous income variable and a logit model based on a categorical binary variable that classifies the population into poor and non-poor statuses.[5] In the remaining of the paper, we refer to the first model as the 'income' model and the second model as the 'poverty' model, with both models leading to poverty predictions. Predicting household poverty with these two models requires three steps which we define as 'Modeling', 'Prediction' and 'Classification' and are described as follows:

Step 1 - Modeling

$$W_i = \alpha + \beta_1 X_i + \eta_i + \epsilon_i \tag{1}$$

$$P_i = \delta + \gamma_1 X_i + \nu_i + \psi_i \tag{2}$$

where i is the unit of observation (usually a household or an individual, household for short), $Wi =$ income, $P_i =$poor where $P_i =1$ if the unit is on or under the poverty line and $P_i = 0$ otherwise, X is a vector of household or individual characteristics, η_i and ν_i are random errors and ϵ_i and ψ_i are model fitting errors.

The second step is the prediction of income or poverty based on the coefficients estimated under the modeling equations:

Step 2 - Prediction

$$\widehat{W_i} = \widehat{\beta_1} X_i + \tilde{\eta_i} + \tilde{\epsilon_i} \tag{3}$$

$$\widehat{P_i} = \widehat{\gamma_1} X_i + \tilde{\nu_i} + \tilde{\psi_i} \tag{4}$$

where $\widehat{W_i}$, $\widehat{P_i}$ are predicted income and poverty and $\tilde{\eta_i}, \tilde{\epsilon_i}, \tilde{\nu_i}, \tilde{\psi_i}$ are the estimated random and model fitting errors. Step 2 is the key step for addressing missing data issues. This is where missing incomes or poverty status are replaced with predicted values.

The third and final step is to divide the population into estimated poor and non-poor groups. For this purpose, the welfare and poverty models critically differ in several important respects. Under the income model, the poverty line is used after the second step to separate the poor from the non-poor. Under the poverty model,

[5]Note that one could use an OLS model with a binary dependent variable but this practice is rare.

the same poverty line is used to separate the poor from the non poor to construct the poor dichotomous variable in step 1 based on observed values. Once the probability of being poor is estimated for missing incomes in Step 2, a probability cut-point is used to separate the poor from the non poor. Therefore, Step 3 can be described as follows:

Step 3 - Classification

$$\begin{aligned} if\ \widehat{W_i} \leq z : i = poor \\ else:\ i = nonpoor \end{aligned} \tag{5}$$

$$\begin{aligned} if\ \widehat{P_i} > prob* : i = poor \\ else:\ i = nonpoor \end{aligned} \tag{6}$$

where z is the poverty line with $W_{min} \leq z \leq W_{max}$ and $prob*$ is a probability cut-point with $0 \leq p \leq 1$ that can be arbitrary or defined with some form of optimization criteria.

A second important difference between the welfare and poverty models is that the income model is typically estimated with an Ordinary Least Squares (OLS) estimator whereas the poverty model is estimated with a Logit or Probit maximum likelihood estimator.

A third difference is that the income model produces income predictions whereas the poverty model produces probabilities of poverty predictions. One can easily turn the monetary predictions from the income model into probability of poverty predictions. In fact, for each poverty line $z = x_0, ..., x_n$ the probability of poverty of a household with income x is 1-F(x). Therefore, we can express both models in terms of probabilities of being poor. However, in practice, scholars have used income or probability of poverty predictions depending on the model used. This implies that comparisons between the two models can only be made after the classification step.

Poverty predictions from both models can be improved after Step 2. The OLS income model produces a distribution of predicted values that is narrower than the true distribution. This is a statistical artifact that has important implications for poverty predictions and that has induced scholars working on cross-survey imputations to propose specific solutions. Poverty predictions from the logit/probit models can also be improved by shifting the probability threshold in order to optimize the trade-off between coverage and leakage. This is done using Receiver-Operating Characteristic (ROC) curves and indexes initially introduced in clinical medicines but also used by poverty specialists (Wodon, 1997; Verme and Gigliarano, 2019). These adjustments

will be considered further in the paper.

The income and poverty models described above is what we refer to as 'classic econometrics' models. We also consider three families of machine learning models: Decision Trees, Regularization, and Neural Networks. In particular, we use Random Forest, Elastic Nets, and Neural Networks with two hidden layers as representative choices of these families. As already discussed, these models are the most popular among economists. Regularization models rely on the same OLS and Logistic models described with the important difference of 'regularization' as a method to shrink parameters. Random forest uses its own classification method based on entropy measures used to split the data in groups as homogeneous as possible and a random selection process for data and variables (bootstrap aggregation) to obtain optimal out-of-sample predictions. Neural networks can be seen as parametric functions such as OLS models with a very high number of parameters that are determined by the trial and error process in-built in the model.[6]

4.2 Confusion matrix and Type I and II errors

All poverty prediction models illustrated above will result in true and false predictions that are best illustrated with a confusion matrix (also known as error matrix or contingency table with two entries) resulting after Step 3 of the modeling exercise (Table1). The matrix divides the population into four groups based on whether predictions are correct or not: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). The primary objective of any classification exercise is to maximize TP and TN and minimize FP and FN. Incorrect classifications result in errors Type I and Type II. In the case of poverty predictions, Type I error refers to non-poor persons who are erroneously predicted as being poor. This error is also known as False Positive Rate (FPR), inclusion error or leakage rate and is defined as FP/(FP+TN). Type II error refers to persons who are poor but are erroneously predicted to be non-poor. This error is also known as False Negative Rate (FNR), exclusion error or undercoverage rate and is defined as FN/(FN+TP).

4.3 Objective functions

To determine the objective function to optimize when predicting poverty, one has to be clear about the type of error to minimize. Both Type I and Type II errors can be

[6] All the machine learning models used by this paper are explained in details in Appendix 1 showing algebraically and in simple words the differences of these models from classic econometric models.

Table 1: True and Predicted Poverty Confusion Matrix

		Predicted Poverty	
		Non-Poor = 0	**Poor = 1**
True Poverty	**Non-poor = 0**	True Negative (TN) [1,1]	False Positive (FP) [1,2]
	Poor = 1	False Negative (FN) [2,1]	True Positive (TP) [2,2]

Note: [x,y] indicates row and column.

regarded as important from the perspective of an administrator of a poverty reduction program. Minimizing Type II error (exclusion error) is clearly more important from a poverty perspective but Type I error (inclusion error) may also be considered important if budgets are constrained, which is a common feature of poverty reduction programs worldwide. How much importance should be given to each objective is, of course, a matter of *preferences* and the trade-offs between the two objectives also depend on the relative *cost* of inclusion or exclusion, which is case/country specific.

Also, in the case of poverty predictions, the objective function to consider is different depending on whether one is interested in estimating the poverty rate as a population statistics (anonymous case), or estimating the poverty status correctly for each household (or individual, non-anonymous case). Below, we consider these two cases in turn.

Population poverty. If the objective is to predict the poverty rate for the population, it is not essential to minimize both Type I and Type II errors. It is sufficient to minimize the difference between the true population poverty rate P and the predicted poverty rate $\widetilde{P}$:

$$min(P - \widetilde{P}) = min[(\widehat{P} + \epsilon) - \widehat{P}] = min(\epsilon) \tag{7}$$

If we refer to the confusion matrix, this is equivalent to maximizing the sum of the true predictions ($max(TN + TP)$) or minimizing the sum of the false predictions ($min(FN + FP)$) irrespective of the actual TN or TP (FN or FP) values. In econometric terms, this amounts to minimizing the average *model* error term for the population (not the *idiosyncratic* error term which averages zero). In this case, and provided we are conducting an experiment where we know the true poverty rate, a possible test to evaluate the performance of the models is a means difference test between the true and predicted poverty rates. However, a means difference test between true and predicted poverty can only be conducted in an experimental context where the true poverty rate is known. In statistics, prediction errors are usually

evaluated with a range of indicators such as the Mean Bias Error (MBE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), the fraction of prediction within a Factor of Two (FACT2), correlation coefficient (R), or Index of Agreement (IA). These indicators are also been used to validate poverty predictions when the true poverty rate is unknown.

Household (individual) poverty. Estimating the correct population poverty rate may not be sufficient if one has an interest in correctly estimating the poverty status of each household included in the sample. This complicates the objective function and the optimization process as we now have two elements to maximize (TN and TP) or minimize (FP and FN). We also need to attribute relative preferences to the two elements unless we consider the two elements equally valuable. One simple way to do that is to maximize the weighted sum of TP and TN as

$$max[a * TP + b * TN]. \tag{8}$$

With a and b indicating preferences for TP and TN. In general, one would prefer to maximize TP and maximize coverage as opposed to maximizing TN and minimizing leakage. However, budget considerations may also be important and different policy makers may have different preferences for a and b.

Maximizing the weighted sum of TP and TN may also not be the best alternative to evaluate the performance of a model for targeting. In addition to the FPR (Type I) and the FNR (Type II) ratios, other popular ratios are the True Positive Rate, sensitivity or recall (TPR=TP/(FN+TP)), the True Negative Rate or specificity (TNR=TN/(TN+FP)), precision (TP/(TP+FP)) or the False Discovery Rate (FP/(TP+FP), the accuracy ratio ($(TP + TN)/N$), Pearson's Chi squared, and F2 statistics ($5 * TP/(5 * TP + 4 * FN * FP)$). All these objective functions are constructed starting from the same confusion matrix. The difference between these functions is simply the weight attributed to each of the four elements in the matrix. In a sense, they are different ways of expressing preferences for different types of errors. Throughout the paper, we will focus on $max(TP + TN)$ and $max[a * TP + b * TN]$ but we will also use these others functions to illustrate how different preferences may lead to different choice of prediction model.

5 Data

In order to observe the true poverty rate and measure the true prediction error, we generate a dummy data set from real data characterized by an extremely low non-response rate, and then corrupt the data with alternative missing data patterns.

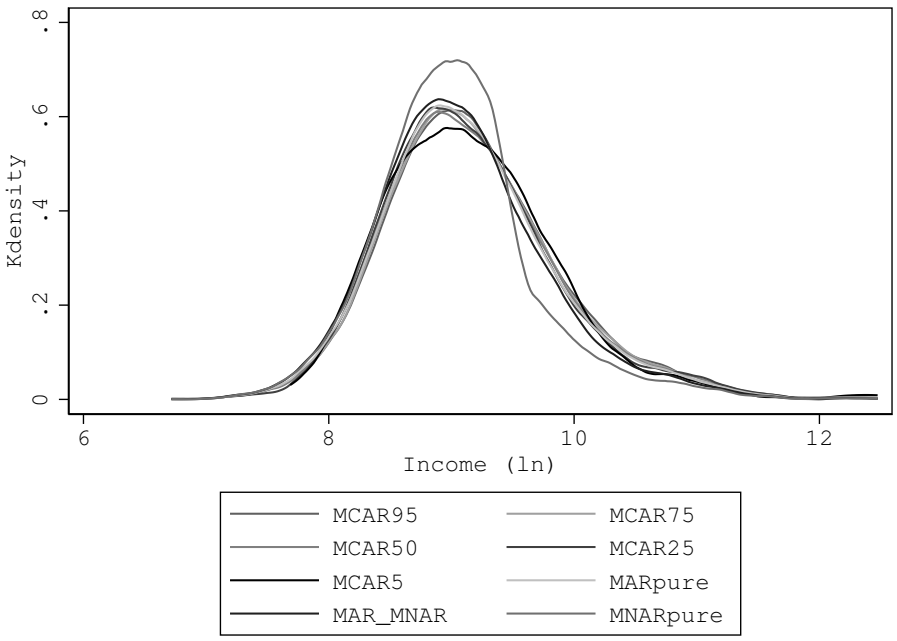
We take the 2007 household consumption survey of Morocco, which is a data set of a middle income country known for its good quality and a non-response rate below 2%, an exceptionally low value by global standards. The sample we use is the one publicly provided by the statistical agency of Morocco which includes 7,062 household observations. The main variable of interest is income per capita. The main statistics of interest is the poverty rate calculated with income per capita on households rather than individuals for simplicity to avoid a discussion on individual weights, which is beyond the scope of this paper. The poverty rate is therefore the share of poor households in the population of households. Summary statistics of all variables considered in this paper are reported in Annex with income reported in local currency. None of the poverty statistics reported in this paper should be considered as an accurate estimate of poverty for Morocco.[7].

In order to compare models' performance when missing data patterns change, we corrupt the initial complete data set mimicking eight missing data patterns: five MCAR selecting randomly different shares of missing data (5, 25, 50, 75, and 95%), "MAR pure" meaning that we randomly selected 50% of the sample conditional on one independent variable that is not correlated with income (we use working individuals in the secondary sector which we tested for independence of income), "MAR-MNAR" where we randomly selected 50% of the sample conditional on a variable which is associated with income (we use household size¡5), and "MNARpure" where we randomly selected 50% of the sample conditional on income only (we use income¿mean income). The most common and relevant case for this paper is MAR-MNAR whereas MNARpure is a rare case in empirical contexts.

The resulting income distributions are plotted in Figure 1. It is shown that the distributions are different in size and shape. Comparing models' performance across the eight data set generated is, therefore, a test across different missing data patterns, but also a test across different shapes of the income distribution as if we were comparing different data sets.

[7]Full information on the survey can be obtained from the High Commission for the Plan of Morocco (https://www.hcp.ma.) and from Douidich et al. (2016)

Figure 1: Distributions of Income with Missing Data Patterns



6 Predicting poverty with econometric and machine learning models

6.1 Baseline econometric models

A simple example may illustrate some of the challenges associated with the estimation of poverty with predicted incomes. Using the complete data set with no missing incomes, we compare poverty estimates based on predicted values derived from the observed full distribution of incomes, which implies that all predictions will be in-sample. In other words, we predict all incomes based on the fully observed distribution of incomes. This is useful to understand the statistical implications of predicting incomes with OLS and Logit models when all information on incomes is available. In practice and for practitioners, this is not a recurrent case since one can use original observations rather than predicted values to estimate poverty. But in some cases, such as for vulnerability analyses or cross-survey imputations, one may want to use predicted values from the fully observed distribution of incomes to estimate the probability of future poverty (vulnerability) or the poverty rate in a different sample (cross-survey imputations).

We compare four models: 1) OLS with continuous dependent variable and predicted incomes; 2) OLS with continuous dependent variable and predicted probabilities of poverty; 3) OLS with binary dependent variable and predicted probabilities of poverty and 4) Logit with binary dependent variable and predicted probabilities of poverty. In all these cases, the final objective is to estimate poverty after classification. The difference between these models is whether we estimate poverty based on predicted incomes applying the poverty line on the post-predictions distribution (1), estimate the probability of poverty after the income model and use a probability cut-off point to determine poor/non-poor status (2) or estimate the probability of poverty after an OLS or logit model with a poverty status dichotomous dependent variable and then use a probability cut-off point to separate the poor from the non-poor (4). Preferences for each of these models vary across practitioners but all these four models have been used in published journals' articles to estimate poverty with predicted values (Ravallion 1996; Gibson 2019). The continuous dependent variable and the poverty lines are set in logs so that we avoid the issue of converting logs of income back into incomes from the log-linear model (Smearing transformation). We test results with three poverty lines set at the 25th, 50th (median) and 75th percentile.

Table 2 shows the results. With a low poverty line (25th percentile) all models underestimate poverty significantly whereas a high poverty line (75th percentile)

results in poverty over-estimations across models. Poverty estimates are closer to the true values with the poverty line set at the median value, although some of the models severely under-estimate poverty also in this case.[8]

Table 2: Models' Comparison

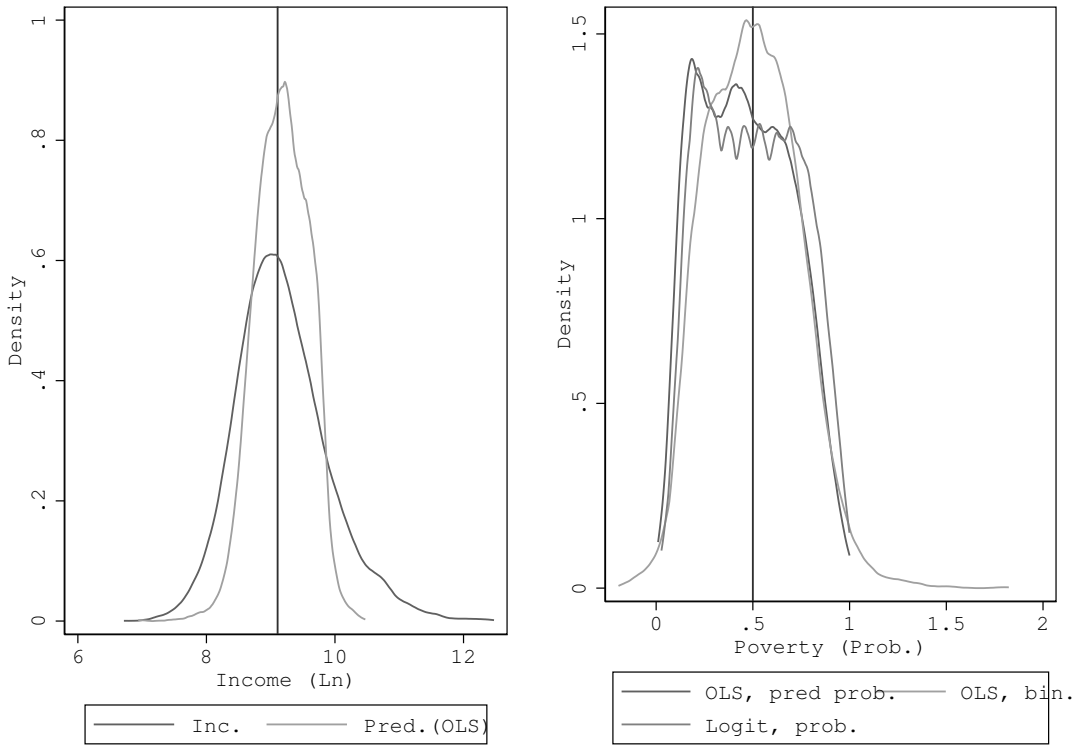
	pl25	pl50	pl75
OLSpredinc	12.36	43.17	82.21
stderr	.39	.59	.46
OLSpredprob	12.36	43.17	82.21
stderr	.39	.59	.46
OLSbinary	8.24	48.58	88.23
stderr	.33	.59	.38
Logitprob	11.51	49.22	84.18
stderr	.38	.59	.43
OLSbinary	12.45	48.58	83.55
stderr	.39	.59	.44

Figure 2 illustrates some of the issues associated with these predictions. The left-hand panel shows income and predicted incomes in log form from the OLS continuous dependent variable model and the right-hand panel shows the probabilities of poverty derived from the other models considered in Table 2 (poverty line = median value). The continuous case shows how predicted incomes are a much narrower distribution than incomes resulting in an underestimation of poverty. The discrete case shows how the distributions of predicted probabilities are very different across the three models considered particularly around the standard probability threshold used to divide poor and non-poor people (0.5). This results in rather different predictions of the poverty rate.

Living aside questions of measurement error that we described as normally distributed with zero mean and of sampling errors, these biases are explained by two main factors. The first factor is purely statistical and is the fact that predictions in the continuous case result in a narrower distribution of incomes as compared to the original distribution. This implies that the tails of the predicted income distribution are shallow as compared to the original distribution. This is well-known of course but what is important to keep in mind is the effect on poverty with severe under and over

[8]Standard errors in Table 2 are estimated with the standard formula for proportions. However, it should be noted that this is likely to underestimate the standard error when one estimates poverty based on probabilities of being poor as in models 2) and 4).

Figure 2: Predicted Values and Probabilities



estimations with low or high poverty lines respectively. A low poverty line results in underestimations because the left tail of the distribution has less observations than it should (there are less poor than there should be). Vice-versa, a high poverty line results in over estimation of poverty because the right tail of the distribution has less observations than it should (there are less non-poor than there should be).

The second factor depends on the explanatory power of the original model. In our case, the first OLS model has an explanatory power (R squared) of 33.2%. This results in many predictions that are incorrectly classified as poor/non-poor contributing to the overall bias. However, the direction of this bias cannot be anticipated unless we know for certain that the original set of independent variables excludes variables that matter for the poor or the non-poor only, therefore systematically under or overestimating poverty. In general, one cannot anticipate the direction of the bias due to model specification implying that, for poverty estimations, the statistical

factor should be expected to be the prevalent source of bias for poverty estimations in models with a continuous dependent variable model.

This simple statistical artifact is often ignored when researchers estimate statistics from partially or totally predicted values because many of the statistics of interest are means and the OLS model centers predictions around the mean by design. However, when one estimates percentages or proportions that are far from the mean, as it is often the case for poverty measures, OLS predictions are no longer suitable for accurate predictions. The cross-survey imputation literature solves this problem using spatial variability to reconstitute a distribution of predicted values that mimics in shape the original distribution. But many practitioners that use OLS imputations to replace missing values are most likely to underestimate poverty.

This section has shown that, even with all incomes observed, one cannot exclude biased predictions of poverty because no model predicts incomes perfectly, and predicting incomes with OLS or Logit models will always result in bias distributions of incomes on the tails, which affects the estimation of poverty.

6.2 Classic econometric Vs. machine learning models

We now expand comparisons to machine learning models using the same complete set of observations as before and predictions of all incomes[9], and compare models with a full set of objective functions that can be used to compare the performance of prediction models. In particular, we compare the performance of eight models: The welfare and poverty models which we described as classic econometric models, and random forest, elastic net and neural network models which we selected as representative models among machine learning models. All models will be estimated in two flavors, with continuous and categorical dependent variable. We label these models wcn, rcn, ecn, ncn, pct, rct, ect and nct where 'w' stands for welfare, 'p' for poverty, 'r' for random forest, 'e' for elastic net, 'n' for neural network, 'cn' for continuous and 'ct' for categorical model. This allows us comparing the performance of econometric and machine learning models and also the performance of continuous and dichotomous dependent variable models. Note that these are "naive" comparisons as we are using off the shelves Stata packages with none of the models being optimized or tuned, something that we will address later in the paper.

For all models, we use the same poverty line set at median income and the same set of explanatory variables with no interactions between variables[10], we do

[9] As in the previous section, all predictions are, therefore, in-sample.

[10] Some models such as random forest will work in a way that amounts to interacting variables but this is not done by design with the inclusion of interactions variables among regressors.

not use any kind of weight and we do not use clustering of standard errors or any other estimation options. All estimations are conducted in Stata[11] and all codes are available on request. For the ML models, we use the simplest possible specification as allowed by the Stata codes used in this paper. This is a naive choice of course because the tuning of ML models is what makes them effective, but the optimization of these models require normative decisions that may vary across specialists, and poverty analysts are not necessarily ML specialists. In other words, this choice allows us to take the specialists' abilities to optimize models out of the picture and also capture basic applications that non-specialists are likely to use. We leave the questions of tuning and optimization to the next sections. In this section, we should keep in mind that ML models are expected to under-perform.

Table 3 compares these baseline models. The top of the table reports the true poverty rate set at 50% by design (poverty lines across the income distribution are tested in the next section), predicted poverty rates, the difference and the t-tests for means difference between the true and predicted poverty rates. We also report the share of true positives and negatives which can be interpreted as a simple objective function with type I and II errors given the same weight and two alternative functions where we give larger preference for TP and TN respectively (prefTN with $a = 1.25$ and $b = 0.75$; prefTP with $a = 0.25$ and $b = 1.25$). The rest of the indicators reported are those which are popular across the social sciences and illustrated under the objective function section. Each indicator gives different weights to the different cells of the confusion matrix. With the exception of the leakage and undercoverage rates, higher values indicate better performance.

The table shows an overall better performance of dichotomous dependent variable models as compared to continuous models, and a better performance of the random forest model as compared to all other models. This is clearly visible when poverty rate predictions and shares of true positive and negative predicted values are compared meaning that this superiority persists if we consider anonymous population poverty and non-anonymous household poverty. These findings also persist if we consider Chi2, Chi2r and F2 evaluation functions.[12] Results may differ, instead, if we consider indicators that privilege certain cells of the confusion matrix such sensitivity vs. specificity, leakage rate vs. undercoverage rate, or precision vs. accuracy. Once we

[11] The Stata commands used for the estimation of the different models are: 'reg' (income model), 'logit' (poverty model), rforest (random forest), 'elasticnet' (elastic net, continuous and dichotomous), and 'mpl2' (Neural Network).

[12] The Chi tests measure the expected distribution of incomes against the actual distribution of incomes. The F measure is the harmonic mean of precision and recall with F2 adding a parameter that allows for more or less weight attributed to precision or recall. See section on confusion matrix for the F2 formula we use.

introduce preferences for certain outcomes, the choice of optimal model may change.

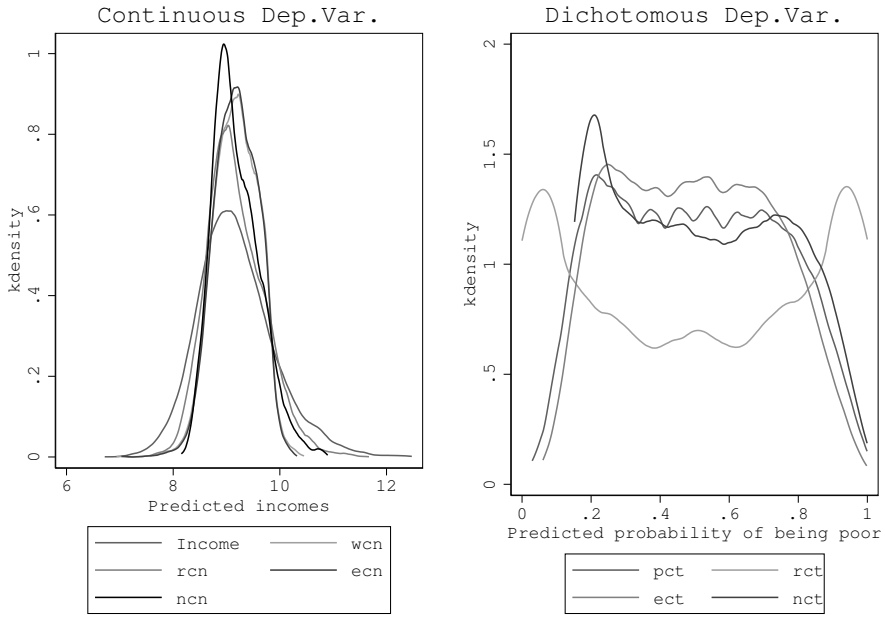
Table 3: Poverty Predictions

	wcn	rcn	ecn	ncn	pct	rct	ect	nct
TruePov	50	50	50	50	50	50	50	50
PredPov	43.17	50.08	43.2	54.77	49.22	49.97	49.39	50.64
Diff.	6.83	-.08	6.8	-4.77	.78	.03	.61	-.64
Diff.(tstat)	10.49	-.17	10.46	-7.54	1.22	.07	.95	-1
TrueShare	69.67	83.12	69.73	71.52	71.01	87.68	70.82	71.1
PrefTP	67.96	83.14	68.03	72.72	70.82	87.67	70.66	71.26
PrefTN	71.37	83.1	71.42	70.33	71.21	87.69	70.97	70.94
Observations	7062	7062	7062	7062	7062	7062	7062	7062
TN	2701	2932	2702	2357	2535	3097	2522	2488
FP	830	599	829	1174	996	434	1009	1043
FN	1312	593	1309	837	1051	436	1052	998
TP	2219	2938	2222	2694	2480	3095	2479	2533
TPR=CR=Sens.	62.84	83.21	62.93	76.3	70.24	87.65	70.21	71.74
TNR=Spec.	76.49	83.04	76.52	66.75	71.79	87.71	71.42	70.46
FPR=IE=LR	23.51	16.96	23.48	33.25	28.21	12.29	28.58	29.54
FNR=ER=UR	37.16	16.79	37.07	23.7	29.76	12.35	29.79	28.26
chi2	1113.54	3098.8	1119.79	1320.67	1247.69	4010.72	1224.14	1257.7
chi2lr	1147.57	3378.25	1154.19	1367.2	1287.34	4518.34	1262.24	1298
Precision	72.78	83.06	72.83	69.65	71.35	87.7	71.07	70.83
Accuracy	69.67	83.12	69.73	71.52	71.01	87.68	70.82	71.1
F2	64.61	83.18	64.69	74.87	70.45	87.66	70.38	71.55

To better understand what determines these findings, it is useful to plot the distributions of predicted values in the continuous case and the distributions of the predicted probabilities in the dichotomous case. These can be found in Figures 3.

Figure 3 shows that the different models perform relatively well in some parts of the distributions but not in others. For the continuous case, all models tend to better perform around the center of the distribution and much less on the tails with the OLS and Elastic Net models being particularly poor on the tails and the Neural Network model being better on the lower tail but very poor on the upper tail. The only model that performs well around the middle of the distribution and has a relatively better performance on the tails is the random forest model. For the dichotomous case, we see similar patterns for all models except the random forest model. This model seems to be better able to split the poor and non-poor into two clearly distinguishable groups as opposed to the other models which have a large area of predictions that could easily switch between poor and non-poor depending on the probability cut-point chosen (an arbitrary choice of 50% in our case. Other

Figure 3: Categorical Dep. Var. Models



thresholds including optimized thresholds are considered in the next section). Again, the random forest model would seem less susceptible to changes in the distribution of incomes and also to the probability cut point chosen to split the poor and the non-poor.

In essence, with a 50% poverty line and by predicting the full distribution of incomes, dichotomous dependent variable models perform better than continuous models but they can be sensitive to the choice of probability cut point used for classification. The random forest model seems less sensitive than other models to changes in these parameters.

7 Robustness and sensitivity tests

7.1 Missing data patterns

Most empirical estimations of poverty have to address some form of missing data issue. To see how different prediction models perform with different forms of missing data we compare the performance of the baseline econometric and ML models using

the corrupted samples illustrated in the data section. Table 4 compares estimated poverty rates across the different models and missing data patterns.[13] Note that, by imposing different shares and patterns of missing observations, we are also dictating the partition of observations between in-sample (observed incomes used for training the model) and out-of-sample (unobserved incomes used for testing the model). We are also testing, therefore, how models behave when the partition between training and testing samples changes. This is particularly relevant for machine learning models.[14]

We can see that all models struggle to maintain accuracy as the share of missing observations increases from 5 to 95% with the exception of random forest in the continuous case and, to a lesser extent, in the dichotomous case. With only 5% of missing observations, all models perform rather well but beyond that threshold income models tend to increasingly underestimate poverty whereas poverty models tend to increasingly overestimate poverty. With MARpure, all models seem to perform relatively well with the exception of the welfare OLS and elastic net models. With MAR-MNAR, all models perform poorly with the exception of random forest. With MNARpure, none of the models performs well. We clearly see that random forest handles various types of missing observations shares and patterns better than other models, although this model too struggles with MNARpure. Considering that most surveys have a share of missing observations higher than 5%, missing observations should always be of concern, even if they are random. MNARpure is a rare case in practice but MAR-MNAR data are the norm with income variables and only random forest seems to handle this case well.[15]

For the welfare and poverty models, we also report results using multiple imputations as opposed to single imputations. We can see that results do not vary for any of the two cases. This is well-known of course but useful as a reminder. Values predicted with multiple imputation are means across repeated samples with replacement and these means center around the simple mean obtained with single imputation. What multiple imputation does is to improve on the estimate of the standard error, which can be larger or smaller than the standard error obtained with single imputation depending on the specification of the model. Therefore, multiple imputation, *per se*,

[13] Standard errors are omitted for simplicity.

[14] There are more sophisticated methods to partition the training and test samples such as "upsampling". Given the variety of shares and patterns of missing observations tested in this section, we will not discuss or use alternative methods.

[15] It is important to stress here that ML models are not adapted to the size of training sample. Some models like neural network require a minimum size for the training set whereas other models need to adapt the choice of parameters to the training sample size. These aspects are ignored here and we should consider that ML models can be improved as shown further in the paper.

does not improve on mean estimates such as the poverty rate, it can only improve on the standard error and the confidence interval, i.e. the confidence we have in the mean estimate.

Table 4: Type and Share of Missing Observations (Pov.Line=Median)

	wcn	wcnMI	rcn	ecn	ncn	pct	pctMI	rct	ect	nct
MCAR95	49.6	49.6	50	49.6	49.1	49.9	49.9	50.1	49.9	49.8
MCAR75	47.8	47.8	49.4	47.8	51.3	49.6	49.6	49.6	49.5	48.6
MCAR50	45.7	45.7	48.7	45.5	59.4	49.2	49.2	49.5	49.7	49.5
MCAR25	43.6	43.6	47.4	43.6	12.5	49.3	49.3	50.2	49.1	48.5
MCAR5	44.3	44.3	50.5	44.9	58.3	52.1	52.1	51.6	51.8	43.8
MARpure	44.4	44.4	47.8	44.3	29.1	47.2	47.2	49.1	47	48.9
MAR_MNAR	42.3	42.3	43.3	42.2	40.4	43.1	43.1	44.4	43.2	43.9
MNARpure	44.4	44.4	45.7	44.4	46.1	44.5	44.5	46.2	44.5	44.4

Note: Numbers such as "95" represent the share of observed incomes.

7.2 Poverty lines

As we have already seen, some models can perform better to predict poverty around certain parts of the income distribution, which means that shifting the poverty line along the distribution may result in different relative performance of the different models. In this section, we consider five poverty lines set at 5, 25, 50, 75 and 95 percentiles of the income distribution to compare outcomes of the different models across these choices. We then provide a stochastic dominance analysis of first degree by comparing the Cumulative Distribution Functions (CDFs) of predicted incomes and poverty in the case of the continuous and dichotomous dependent variables models.

Table 5 provides results for changes in the poverty line. As in the first comparison table, we are predicting incomes for the entire distribution, which makes results extreme. We can see that all models perform better when the poverty line is close to the median value of the income distribution and, as already discussed, dichotomous dependent variable models perform better than continuous dependent variable models while random forest seems to outperform other models. As we move away from the median value, all models start to struggle and with poverty lines around the 25th or 75th percentile poverty predictions are already very much off the mark with only random forest coming anywhere close to the true poverty rate. This is remarkable considering that poverty lines are most often in the range of the 20th-40th percentile

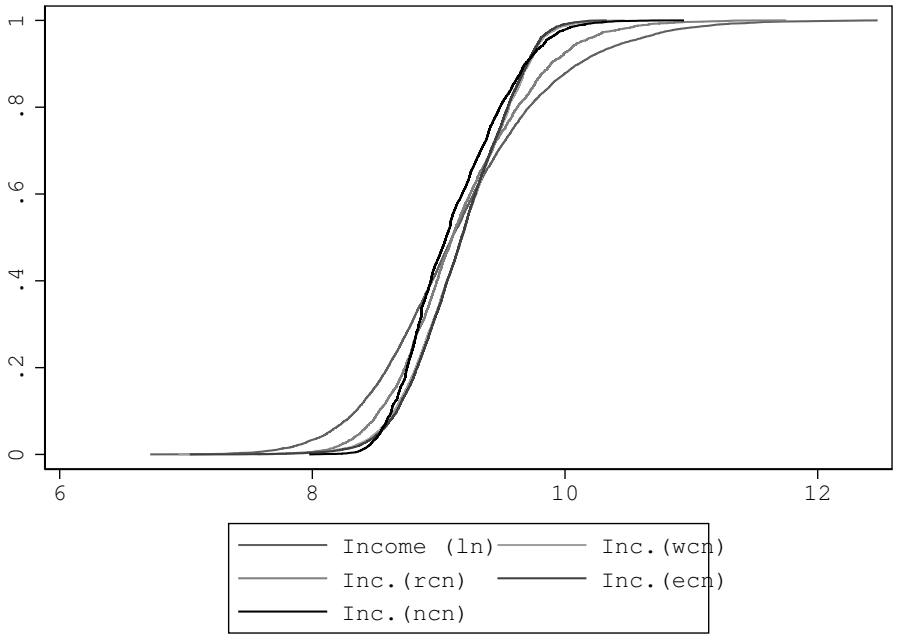
in middle and low-income countries with rates above 40% only seen in extremely poor countries. Recall that scholars working on vulnerability or cross-survey imputations do predict the full distribution of incomes. In wealthy countries, non-response rates may also be as high as 50% so that predicting half of incomes is a possibility.

Table 5: Models' Comparison: Poverty Lines and Distributions

	wcn	rcn	ecn	ncn	pct	rct	ect	nct
PL5	.8	1.1	.6	.1	.2	3.3	.2	0
PL25	12.4	17.4	11.4	14.9	11.5	21.8	9.5	14.3
PL40	29.9	36.5	29.3	41.1	33.6	38.8	32	36.1
PL50	43.2	50.1	43.2	54.8	49.2	50	49.4	50.6
PL60	58.6	61.8	59	66.3	64.1	61.7	65.1	65.5
PL75	82.2	78.5	83	85	84.2	77.3	86.1	82.8
PL95	100	100	100	99.9	99.4	96.9	99.9	100

Looking at the Cumulative Distribution Functions of predicted values help us understanding the results (Figure 4). It is clear that CDFs cut across each other and some also cut across the original income distribution (left-hand panel). This means that there is no absolute dominance along the distribution with some models predicting income or poverty consistently lower or higher than other models or the original income distribution. Models that may perform better with low poverty lines (high probability thresholds) may not perform better with high poverty lines (low probability thresholds) and vice-versa. Predicted values are the further away from the original distribution on the tails explaining why poverty predictions are far from the true values when we use poverty lines that approach the 25th or 75th percentile. The real problem with predictions of any model really lies on the incapacity of these models to fit the tails of the true income distribution. Random forest is an exception in this respect as the CDF of predicted values is the closest of all models to the tails of the true income distribution in the continuous case. In the dichotomous case, random forest shows the most extreme curve on the tails indicating that its mass is concentrated closer to 0 and 1 probabilities. This makes classification of observations easier and more accurate with less observations located around the cutpoint. Improvements to ML models could also be achieved by giving more weight to the tails of the distribution but this assumes that we have some ex-ante knowledge about the distribution of missing incomes, which is rare in practice.

Figure 4: Predicted Incomes and Poverty (CDFs)



7.3 Model specification

The set of independent variables and the explanatory power that this set determines may also be a discriminatory factor for the choice of an optimal prediction model. If an important variable is not included into the prediction equation, none of the models will benefit from this variable. However, machine learning models have the ability to improve on the use of the existing set of independent variables by including or omitting variables, or interacting them. Therefore, the initial set of independent variables may benefit some models more than others. We will see that for certain models this applies not only to the number of independent variables but also to the order in which these variables are listed in the model.

In Table 6 we compare the initial set of independent variables we used thus far (Model1) with three other sets (Models2, Model3 and Model4). Model2 uses the same variables as model1 but changes the order by placing the continuous dependent variables at the bottom rather than at the top. Model 3 is a reduced model that includes all original variables as in Model1 except age and hhsize, two important predictors in our model. This reduces the R2 of the welfare OLS model from 0.33 to 0.22 and the Pseudo-R2 of the Logit poverty model from 0.18 to 0.11. Model4 is an even smaller model that keeps only the variables male, marital status and urban from Model1, which reduces the R2 of the OLS model to 0.12 and the Pseudo-R2 of the Logit model to 0.06. We can therefore compare four models with different orders of the independent variables and different sets of explanatory variables.

A first obvious result is that, with the largest model (Model1), all models tend to perform better whereas none of the models performs well with the most reduced of the models (Model4). However, the order in which the independent variables are listed is important for the neural network model and, to a minor extent, the random forest model. This is a question that arises with ML models that rely on randomized processes such as RF and NN.[16] We will return to this question in the next section of the paper. With the most parsimonious of the models (Model4), differences in poverty estimations across models almost disappear. With only three variables that explain little of the outcome variance, there is not much that machine learning models can add to simple OLS or logit models unless the model that best fits the data is not a linear model.

An additional observation is that with an intermediary set of variables (Model3), dichotomous models still perform better than continuous models, the random forest model holds up well among the continuous models, and the poverty logit model

[16]For a discussion of permutation invariance in neural network models see: https://gmarti.gitlab.io/ml/2019/09/01/correl-invariance-permutations-nn.html.

outperforms all other dichotomous models. With parsimonious models, one may want to stick to a simple logit model. However, ML models can also be optimized for models with few predictors, which may improve further on the observed performance of the ML models we considered.

Table 6: Models' Comparison: Model Specification

	wcn	rcn	ecn	ncn	pct	rct	ect	nct
Model1	43.2	50.1	43.2	48	49.2	50	49.4	50.6
Model2	43.2	50.1	43.2	43	49.2	50.2	49.4	54.5
Model3	42.3	51.9	43	45.4	50.7	53.5	51.1	52.5
Model4	39.6	39.6	39.6	38.4	39.6	39.6	39.6	41.8

8 Models' calibration

So far, we have been using all models with the most basic specifications and tuning as provided by the Stata packages we used. In this section, we test how varying models' parameters can affect the estimation of poverty. With the exception of weighting discussed below, these calibrations are model specific.

8.1 Weighting

We have seen the inability of certain models to make proper income predictions in the tails of a distribution. One way of addressing this problem would be to attribute more weight to observations located in the tails. Some of the literature on top incomes has followed this strategy by attributing to observed incomes a weight equal to the inverse of the probability of incomes being observed (estimated at the local level such as Probability Sampling Units (PSU), districts or regions) and this approach has been shown to be very effective in improving estimates of inequality across countries (Korinek et al., 2007; Hlasny et al., 2021). However, this approach requires estimating the probability of income being observed which, in turns, requires knowledge on response rates by geographical areas, information that is not always available. In all other cases when no information is available on the distribution of missing incomes, weighting amounts to a tentative experiment which is hard to validate.

8.2 OLS Regression

The real shortcoming of a linear regression model is its inability to predict incomes on the tails correctly, a problem that extends to elastic net models. This is simply a statistical artifact of OLS models which results in distributions of predicted values that are narrower than the original distributions, particularly if the explanatory power of the model is low. As surprising as it may seem, this is a problem that is routinely ignored in empirical works. One strand of the poverty literature that focused on this problem is the cross-survey imputation literature which proposed to address it by correcting the error term. In essence, the error term can be split into an idiosyncratic error term and a model error term. By estimating the model error term using the original empirical distribution or a theoretical normal distribution, one can add this error back into the predicted values mimicking in this way the variance of the original distribution. This is what we do in this section replicating the same technique used in cross-survey imputations.

Results are shown in Table 7 providing corrections using empirical and theoretical normal data for a range of poverty lines. This form of imputation improves results substantially for all poverty lines bringing estimations much closer to the true values. Imputation with empirical data performs better than the one with normal data but both methods improve estimations very visibly. In the presence of missing data, particularly when the share of missing data is very large, and when the true poverty rate is expected to be far from the center of the distribution, it is essential to use this method when estimating poverty with OLS or elastic nets models. This simply confirms a fact that is very well known among cross-survey imputation specialists (Dang et al., 2019).

Table 7: Models' Comparison with and without error adjustments

	PL5	PL25	PL40	PL50	PL60	PL75	PL95
TruePov	5	25	40	50	60	75	95
PovOLS	.7	12.9	31.6	44.4	60.3	83.4	100
st.err.	.1	.6	.8	.8	.8	.6	0
PovImpNorm	7.1	25.6	38.3	47	56.2	72.3	96.7
st.err.	.8	1.8	2.1	2.1	2.1	1.7	.4
PovImpEmp	6.7	25.8	39.3	48.4	57.8	73.6	96.2
st.err.	.8	1.9	2.2	2.2	2.1	1.7	.4

8.3 Logit

It is possible to improve on predictions generated by dichotomous dependent variable models by optimizing the probability cutpoint used to separate the predicted poor from the predicted non-poor when households are classified after the prediction step. As probabilities of being poor vary between 0 and 1, most scholars use a cutpoint of 0.5 for simplicity, which is what we used so far. This is also the threshold that is normally selected if the purpose of the exercise is simply to reproduce the original poverty rate. However, research across the social sciences has shown that one can use Receiver Operating Characteristics (ROC) curves and the Youden index (defined as y=max(sensitivity+specificity-1)), or the max vertical distance between the ROC curve and the chance line) to optimize the cutpoint (see Verme and Gigliarano, 2019 for a detailed discussion).

Table 8 shows poverty rates for the four dichotomous dependent variable models with the optimal cutpoint derived from the model and from artificial cutpoints that we selected between 40 and 60%. The optimal cutpoint is derived from the ROC curve described. An alternative approach to identify the optimal cutpoint would be cross-validation by testing the performance of different cutpoints on out-of-sample predictions. This is grossly mimicked by shifting the cutpoint artificially.

We see that the optimal cutpoint is similar across models except for the random forest model. That is because, as we have already seen, random forest fits the tails of the original income distribution better than the other models and this results in a higher density of predicted poverty towards lower and higher percentiles. This, in turn, results in a higher concavity of the ROC curve and a higher distance of the ROC curve from the 45 degrees line (the chance line). This also explains why random forest overestimates poverty with higher cutpoints and underestimates it with lower cutpoints when compared to other models. The fact that random forest produces thicker tails of predicted probabilities of being poor has very different implications when one optimizes the cutpoint or selects different arbitrary cutpoints. This is particularly important for policy makers working on targeting and interested in non-anonymous predictions. The table also shows that, with the exception of rct,

8.4 Random Forest

In the case of random forest, we have three parameters that are particularly of interest to understand the poverty prediction behavior of this model: the number of iterations (trees), the depth of the trees, and the mtry parameter, which regulates the number of input variables at each iteration. A higher number of trees is preferable but increasing this number increases the computation time. Numbers between 100

Table 8: Poverty rates with different probability cutpoints

	pct	rct	ect	nct
Pov(OptCut)	57.7	33.3	59.8	55.5
OptimalCut	42.5	75.4	42.2	42.7
Pov(Cut=0.40)	62.1	56.7	62.93	60.04
Pov(Cut=0.45)	55.4	54.62	56.24	53.65
Pov(Cut=0.50)	49.2	49.97	49.39	50.64
Pov(Cut=0.55)	43.9	46.11	42.69	42.78
Pov(Cut=0.60)	36.6	43.5	35.67	39.8

and 500 are usually considered good trade-offs. The depth of the trees determines how many data splits should be allowed for variables. A higher depth increases the in-sample prediction capacity but decreases the out-of-sample prediction capacity. The optimal choice may also depend on the type and number of independent variables one has. The mtry parameter can vary from a few variables to the total number of independent variables. Increasing this parameter also increases computation time.

In what follows, we test the model with 1, 5, 10 100 and 1000 for both trees and depth and across a set of poverty lines (5, 25, 40, 50, 60, 75, and 95 percentile).[17]

Table 9 shows results for the number of iterations (trees) and for the continuous and dichotomous dependent variable models. Increasing trees does not seem to improve predictions for neither the continuous nor the dichotomous case with the exception of cases where the poverty line is located around the median of the original income distribution. With extreme poverty lines, increasing iterations does not seem to bring any benefit but it may help when poverty lines are set around center values.

Table 10 repeats the exercise for depth keeping iterations at 100. It shows that shallow trees are incapable of making proper predictions. Only starting from a depth of 10, we observe predictions becoming valuable with poverty lines set around the median value of the income distribution but not with extreme poverty lines. Once we reach a depth of 100, there is no more benefit in increasing depth. In this case, one may want to find the optimal depth as increasing depth is very costly in terms of computational time and may lead to overfitting out-of-sample predictions. Also, increasing trees, while it reduces errors, also reduced the probability of having fully independent trees.

To evaluate the performance of these models in and out-of-sample, we use AUC-

[17]Note that depth is also selected based on the type of model, classification or regression. In classification models, a standard choice is 1 split while 5 splits is standard choice for regressions. Here we apply all splits to all models.

Table 9: Random Forest Iteration (trees) Test

	PL5	PL25	PL40	PL50	PL60	PL75	PL95
rcn1	3.6	23.5	39	49.4	59.8	76.5	96.8
rcn5	1.6	19.3	37.3	50	61.9	78.3	98.1
rcn10	1.4	18.1	37.3	50.1	61.6	78.9	98.1
rcn100	1.1	17.4	36.5	50.1	61.8	78.5	98.2
rcn1000	1.2	17.2	36.3	49.9	61.7	78.5	98.3
rct1	4.2	23.1	37.7	47.9	58.2	74.3	95.3
rct5	3.4	21.8	38.5	50.1	61.2	77.2	96.7
rct10	3.4	21.4	38.4	50.3	60.9	77	96.8
rct100	3.3	21.8	38.8	50	61.7	77.3	96.9
rct1000	3.1	21	39	50.2	61.9	77.7	97

ROC values and plot these values against the level of depth for the train and test sets of observations and with depth varying between 5 and 95. Recall that $AUC = 1$ represents perfect predictions, $AUC = 0$ represents non-overlap between true values and predictions, and $AUC = 0.5$ represents predictions that correspond to random predictions. The train (in-) sample was selected randomly extracting 75% of the original incomes. The test (out-of-) sample is the remaining 25% of observations. We can see in this way how random forest performs in and out-of-sample as we increase the depth of the trees.

Figure 5 shows that the optimal depth which maximizes AUC values is around 6-7 where both the train and test curves are at their max for both the regression and logit models. Around these values, in-sample (train) predictions are better than out-of-sample (test) predictions as we should expect, but the difference is not very large and out-of-sample predictions are very close to in-sample predictions all along the curves. Higher depth reduces in-sample prediction invariably for train and test samples, regression and logit models. Interestingly, when AUC values hit 0.5 (the equivalent of random predictions), there is no more difference between the models, In fact, beyond a depth of 10 (not shown in the graph) out-of-sample prediction outpace in-sample predictions. This simply means that they remain closer to random predictions and they should be ignored as this is not a choice of depth that one would consider.

In order to preserve the consistency of results, it is also important to consider the order of variables as briefly mentioned in the section above. In RF models, changing the order of independent variables can change results, even when seeds are set. This happens when the mtry parameter is set below the total number of independent

Table 10: Random Forest Depth Test

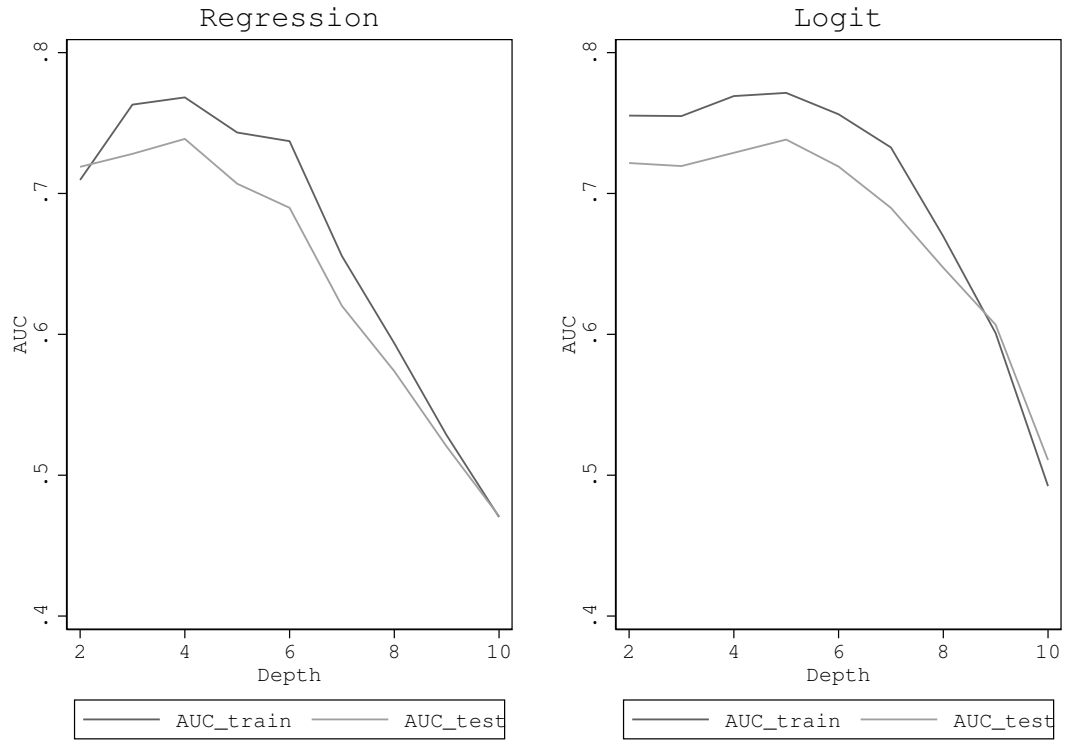
	PL5	PL25	PL40	PL50	PL60	PL75	PL95
rcn1	0	0	0	12.4	99.4	100	100
rcn5	0	0	18.7	43.4	67.3	97.7	100
rcn10	0	6.8	30.1	48.4	64.8	83.3	99.7
rcn100	1.1	17.4	36.5	50.1	61.8	78.5	98.2
rcn1000	1.1	17.4	36.5	50.1	61.8	78.5	98.2
rct1	0	0	0	54	100	100	100
rct5	0	0	22.3	49.8	76.4	98.8	100
rct10	.2	10.3	32.5	51.8	68	86	99.2
rct100	3.3	21.8	38.8	50	61.7	77.3	96.9
rct1000	3.3	21.8	38.8	50	61.7	77.3	96.9

variables and with low number of trees (random sub-space method), or when the explanatory power of the original model is low. That is because RF samples groups of variables in sets chosen based on the position of the independent variables. If the position is changed, different variables are selected (bagged), which results in different trees. To address this issue, one has to make the mtry parameter equal to the total number of independent variables and use a very large number of trees. This strategy will ensure that results are consistent, even if we change the order of independent variables, similarly to OLS or Logit models. This was not the case with the RF model used in this paper because we used one of the standard methods to select mtry (the squared root of the total number of independent variables).

8.5 Elastic Nets

The two parameters to consider in Elastic Nets are the α parameter which controls the relative weight given to the RIDGE and LASSO models (RIDGE with $\alpha = 0$ and LASSO with $\alpha = 1$) and the λ parameter which controls the regularization parameter, with a higher λ resulting in more reduced models (forcing more variables' coefficients to zero. A $\lambda = 0$ is equivalent to an OLS model). Most statistical packages use internal routines to optimize λ based on standard criteria designed to optimize out of sample predictions. When these routines are used, the only important parameter to regulate is α. However, in practice, changing the relative RIDGE/LASSO weight has little effect on predicted means such as the poverty rate and only a marginal effect

Figure 5: ROC-AUC Curves



on standard errors.[18] Imposing a λ, on the other hand, is not recommended in our case as one would want to optimize λ based on out of sample predictions. Therefore, provided we rely on standard routines to select the optimal λ, we don't expect much variance in poverty estimations across elastic nets models. In our particular case, the optimal λ found is also rather close to zero resulting in poverty predictions that are little different from standard OLS models as we have verified throughout the paper. We should also expect this to be often the case with poverty prediction models that have a similar number of variables and similar explanatory power to the model used in this paper. A difference may be observed with initial models that have a very large number of explanatory variables. In these cases, regularization parameters may play

[18]Note that standard errors may be difficult to interpret with Elastic Nets models given the different weights attributed to the two models.

a more important role and result in predictions that depart from a standard OLS model than those we observed. However, the main shortcoming of the OLS model, which is its incapacity to model the tails of the original income distribution, persists with any elastic net model. Focusing on this question is more important than searching for optimal α or λ parameters.

8.6 Neural Network

Neural network models can be very complex and there are tens of parameters that can potentially be set by the user depending on the routine available. As an example, the neural network model used in the paper is mlp2 programmed in Stata and the parameters used are the default parameters provided by this program. These include two hidden layers and the following choices: 1) number of neurons per layer = number of layers of the outcome variable for both layers; 2) Added bias term ; 3) SGD optimizer; 4) Softmax and MSE loss functions for categorical and continuous dependent variables respectively; 5) Variance scaling factor = 1; 6) Maximum attempts to find good initial values of the parameters=10; 7) Learning rate = 0.1; 8) Dropout probability rate = 0; 9) Batch size for training = 50; 10) Maximum number of iterations of optimizer = 100; 11) Choice of metrics is accuracy for softmax and MAE for MSE. Naturally, experimenting with variations in these choices would require a separate paper.

Even if all options listed above are kept constant, neural network models can quickly become unstable and time consuming. A model with one hidden layer and a few neurons is more likely to provide stable results and be fast but also approximates classic linear models and may not be suitable for more complex non-linearities, which is the advantage of using neural networks. Vice-versa, a model with many hidden layers and neurons can become very unstable and very time consuming. It is also worth mentioning that neural network models perform best with very large data sets and big data, which is not the typical case with survey income data.

Stability of results in one important concern that we can test. In Table 11, we run the neural network model multiple times without setting seeds and varying only the number of nodes (neurons) per hidden layer using 20, 40, and 60 nodes for both the continuous and categorical models. We can see that the variance of outcomes across repetitions of the model is quite large and the mean across repetitions of the model does not necessarily converge towards the true value. We repeated this exercise multiple times (not shown in the paper) and we could not find any indication that mean or variance would stabilize, or that the mean would converge towards the true value. With several sets of nodes and at least ten repetitions, this model can

easily run into hours of computation (with the Stata routine that we use, better programming or software may reduce computation time). This feature of neural network models together with their complexity makes these models less appealing for poverty specialists working with empirical data.

Table 11: Neural Networks, Layers and Trials Test

	ncn20	ncn40	ncn60	nct20	nct40	nct60
t1	47.9	51.4	49.9	49.6	55.9	51.3
t2	53.9	51.4	50.1	54.8	54.6	55.1
t3	50.2	51.4	35.7	50.2	54.4	55.6
t4	45.1	47.8	46.1	54.9	51.7	46.2
t5	44.9	48.4	45.8	53.6	61.7	58.7
t6	51.2	53.4	52.9	55.9	54.7	49.4
t7	48.9	55.3	51.1	52.8	51.9	53.6
t8	51.3	51.4	50.6	53.4	53.6	52.5
t9	53	51.9	49.6	51.8	55.4	53.9
t10	0	56.2	49.8	52.4	52.7	50.5
mean	44.6	52.2	48.2	52.9	54.7	52.7
var.	229.6	6.9	21.4	3.7	7.3	11.2

9 Concluding Remarks

The paper has provided a comparative analysis of classic econometric and machine learning models used for the estimation of poverty at the population or household level in the presence of missing data. We carried out an artificial experiment using a dummy data set constructed on real data comparing eight different models and testing the robustness of results to changes in data, parameters and preferences. This strategy allowed us to address two important shortcomings of poverty prediction models: Test how different models perform with different missing data patterns including MNAR, and compare results with the true poverty rate. Below we provide some indications that can help practitioners consider alternative models for poverty predictions. These indications are preliminary, and some may be explained by a naive use of the models, but they can help to orient practitioners in the use of these models and scholars to structure future research.

- Missing observations should always be of concern for poverty predictions unless they are a very low share of observations (say less than 5%). Some models are

not effective in predicting poverty even if missing observations are Missing Completely At random (MCAR).

- Overall and for poverty predictions, no model can be expected to outperform all other models under any circumstance. The paper showed that models' relative performance depends on the original distribution of incomes, the poverty line, models' parameters, the pattern of missing observations, the objective function and policy preferences. No model "dominates" others and predict incomes that are closer to the original distribution of incomes all along the distribution and in all cases tested.

- When missing data are MNAR-pure none of the models studied in this paper performs well. MNAR-pure data are rare and the case of MAR-MNAR data is the most common with empirical income data. In this case, most models struggle to provide good poverty predictions with the exception of random forest.

- The random forest model has proved to be the most consistent in predicting poverty relatively well under almost any condition considered in this paper. This is also consistent with tests conducted on other indicators of deprivation (Andree et al., 2020). It makes this model the most flexible and a preferred candidate when researchers lack key information for making a choice among models such as information about missing data patterns. For these models, it is important to have a sufficiently large number of iterations (trees) to have stable predictions and a proper depth, which may vary from case to case and needs to be tested with out-of-sample predictions.

- Simple OLS models are generally ineffective in predicting poverty accurately if the model error term of predicted values is not adjusted post-estimation and the true poverty rate is distant from the mean. That is because of the narrow distribution of predicted values as compared to the original distribution (a statistical artifact) and the incapacity of these models to predict income on the tails well. This finding extends to elastic nets models which are based on OLS models.

- OLS models can be substantially improved if the model error term is adjusted post-estimation as it is done in the cross-survey imputation literature. The paper found that error adjustments derived from the original empirical distribution are better than those derived from a theoretical normal distribution.

However, both methods are effective and preference for one method over the other may depend on the data at hand.

- Dichotomous dependent variable models tend to perform better, on average, than continuous dependent variable models because they do not suffer from extreme errors on the tails of the distribution of predicted probabilities. This relative superiority should not be given for granted if the model error of OLS models is adjusted to reflect the original variance of the distribution of income.

- Dichotomous dependent variable models can be improved by searching for the optimal probability cutpoint using ROC curves, but only marginally.

- The simple logit model performs, on average, better than machine learning dichotomous dependent variable models with the exception of random forest.

- Elastic net models have a similar performance to OLS and logit models because they use the same functions, but they add layers of complexity that rarely result in better performance than a simple OLS or logit model. The α parameter which regulates the weight of the RIDGE and LASSO components seems to make little difference and does not improve on the fundamental problem of OLS models. The λ parameter which is the regularization parameter is usually optimized by the model itself and this optimization does not seem sufficient to address prediction problems on the tails.

- With ML models re-ordering the variables can lead to different results if these variables are continuous. It is therefore safer to transform these variables into categorical or dummy variables.

- Neural network models are very complex, time consuming and not easy to stabilize. These features make these models not particularly appealing for poverty specialists working on standard poverty analyses and empirical data.

- Multiple imputation is effective in improving on the variance and standard error of estimations but does not affect means. A model that predicts a poverty rate that is far from the true value cannot be fixed by simply using multiple imputation.

- In the case of extremely reduced models with few independent variables (say 2-4 variables), there is not much difference in what model is used. All models will perform equally poorly.

- With a large set of independent variables (say more than 15), machine learning models have a comparative advantage in that they can test alternative reduced models and find the most effective in predicting poverty sparing researchers a complex and time consuming trial and error process.
- Finally, we should expect non-parametric models to handle non-linearities better than parametric linear models. This question is really data specific and will require to be addressed in future research.

To conclude, in the absence of complete information on data, parameters and preferences, and in the absence of a deep understanding of machine learning models, logistic and random forest models should be preferred to OLS and other machine learning models. With time availability and a more nuanced knowledge of machine learning models, a comparative analysis similar to what this paper has provided can help practitioners making better choices.

References

Abelson, B., k. R. Varshney, and J. Sun (2014). Targeting direct cash transfers to the extremely poor. *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, 1563–1572.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*, 864–870.

Aiken, E. L., G. Bedoya, J. E. Blumenstock, and A. Coville (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan. *Journal of Development Economics 161*, 103016.

Andree, B. (2021). Estimating food price inflation from partial surveys. *World Bank Policy Research Working Paper 9886*.

Andree, B., A. Chamorro, A. Kraay, P. Spencer, and D. Wang (2020). Predicting food crises. *World Bank Policy Research Working Paper 9412*.

Atkinson, A., T. Piketty, and E. Saez (2011). Top incomes in the long run of history. *Journal of Economic Literature 49*, 3–71.

Baker, J. L. and M. E. Grosh (1994). Poverty reduction through geographic targeting: How well does it work? *World Development 22*(7), 983–995.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*(626), 1073–1076.

Bollinger, C., B. Hirsch, C. M. Hokayem, and J. Ziliak (2019). Trouble in the tails? what we know about earnings nonresponse 30 years after lillard, smith, and welch. *Journal of Political Economy 127*(5), 2143 – 2185.

Brown, C., M. Ravallion, and D. van de Walle (2018). A poor means test? econometric targeting in africa. *Journal of Development Economics 134*(C), 109–124.

Calvo, C. and S. Dercon (2013). Vulnerability to individual and aggregate poverty. *Social Choice and Welfare 41*(4), 721–740.

Coady, D., M. Grosh, and J. Hoddinott (2004). *Targeting of Transfers in Developing Countries: Review of Lessons and Experience*. The World Bank.

Cowell, F. and M. Victoria-Feser (1996a). Poverty measurement with contaminated data: A robust approach. *European Economic Review 40*, 1761–1771.

Cowell, F. and M.-P. Victoria-Feser (1996b). Robustness properties of inequality measures. *Econometrica 64*, 77–101.

D'Alessio, G. and I. Faiella (2002). Non-response behaviour in the bank of italy's survey of household income and wealth. *Banca D'Italia: Temi di discussione* (462).

D'Alessio, G. and A. Neri (2015). Income and wealth sample estimates consistent with macro aggregates: some experiments. *Banca D'Italia: Questioni di Economia e Finanza, Occasional Papers* (272).

Dang, H., D. Jolliffe, and C. Carletto (2019, July). Data Gaps, Data Incomparability, And Data Imputation: A Review Of Poverty Measurement Methods For Data-Scarce Environments. *Journal of Economic Surveys 33*(3), 757–797.

Dang, H.-A. and P. Verme (2022). Estimating poverty for refugee populations: Can cross-survey imputation methods substitute for data scarcity? *Journal of Population Economics* (forthcoming).

Douidich, M., A. Ezzrari, R. van der Weide, and P. Verme (2016). Estimating quarterly poverty rates using labor force surveys: A primer. *World Bank Economic Review 30*(3), 475–500.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro–level estimation of poverty and inequality. *Econometrica 71*(1), 355–364.

Gibson, J. (2019). Are you estimating the right thing? an editor reflects. *Applied Economic Perspectives and Policy 41*(3), 329–350.

Glewwe, P. (1991). Investigating the determinants of household welfare in cote d'ivoire. *Journal of Development Economics 35*(2), 307–337.

Hlasny, V., L. Ceriani, and P. Verme (2021). Bottom incomes and the measurement of poverty and inequality. *Review of Income and Wealth doi.org/10.1111/roiw.12535*.

Hlasny, V. and P. . Verme (2018a). Top incomes and the measurement of inequality in egypt. *World Bank Economic Review 32*(32), 428–455.

Hlasny, V. and P. . Verme (2021). The impact of top incomes biases on the measurement of inequality in the united states. *Oxford Bulletin of Economics and Statistics* (https://doi.org/10.1111/obes.12472).

Hlasny, V. and P. Verme (2018b). Top incomes and inequality measurement: A comparative analysis of correction methods using the eu silc data. *Econometrics 6*(2), 1–21.

Jean, N., M. Burke, M. Xie, M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Jenkins, S. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica 84*(334), 261–289.

Korinek, A., J. Mistiaen, and M. Ravallion (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality 4*(1), 33–55.

Korinek, A., J. Mistiaen, and M. Ravallion (2007). An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics 136*(1), 213–235.

Lillard, L., J. Smith, and F. Welch (1986). What do we really know about wages? the importance of nonreporting and census imputation. *Journal of Political Economy 94*(3), 489–506.

Mcbride, L. and A. Nichols (2018). Retooling poverty targeting using out-of-sample validation and machine learnin. *World Bank Economic Review 32*(3), 531–550.

Morduch, J. (1994). Poverty and vulnerability. *American Economic Review 84*(2), 221–25.

Ravallion, M. (1996). Issues in measuring and modelling poverty. *Economic Journal 106*(438), 1328–43.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. and R. J. A. Little (2020). *Statistical Analysis with Missing Data, 3rd Edition*. John Wiley and Sons.

Tarozzi, A. and A. Deaton (2009). Using census and survey data to estimate poverty and inequality for small areas. *The Review of Economics and Statistics 91*(4), 773–792.

Verme, P. and C. Gigliarano (2019). Optimal targeting under budget constraints in a humanitarian context. *World Development* (119).

Verme, P., C. Gigliarano, C. Wieser, K. Hedlund, M. Petzoldt, and M. Santacroce (2016). *The Welfare of Syrian Refugees: Evidence from Jordan and Lebanon*. Washington DC: World Bank.

Wodon, Q. (1997). Targeting the poor using roc curves. *World Development 25*(12), 2083–2092.

Annex

Table 12: Data Summary Statistics

Variable	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
household income	7,062	56887	54561	2800	1113157
income per capita	7,062	13118	15098	826	261621
age	7,062	51.64	14.00	15	98
age squared	7,062	2863	1526	225	9604
household size	7,062	5.14	2.43	1	24
male	7,062	0.82	0.38	0	1
marital status	7,062	0.83	0.38	0	1
skills	7,062	0.19	0.39	0	1
urban	7,062	0.60	0.49	0	1
work_salaried	7,062	0.39	0.49	0	1
work_selfemployed	7,062	0.31	0.46	0	1
work_unpaid	7,062	0.00	0.05	0	1
econ.sect._secondary	7,062	0.17	0.37	0	1
econ.sect_tertiary	7,062	0.33	0.47	0	1
out of labor force	7,062	0.26	0.44	0	1