Working Paper Series

Which Model for Poverty Predictions?

Paolo Verme

ECINEQ WP 2020 - 521

ECINEQ 2020 - 521
February 2020

www.ecineq.org

Which Model for Poverty Predictions?

Paolo Verme
World Bank

Abstract

OLS models are the predominant choice for poverty predictions in a variety of contexts such as proxy-means tests, poverty mapping or cross-survey imputations. This paper compares the performance of econometric and machine learning models in predicting poverty using alternative objective functions and stochastic dominance analysis based on coverage curves. It finds that the choice of an optimal model largely depends on the distribution of incomes and the poverty line. Comparing the performance of different econometric and machine learning models is therefore an important step in the process of optimizing poverty predictions and targeting ratios.

Keywords: Welfare Modelling; Income Distributions; Poverty Predictions; Imputations.

JEL Classification: D31, D63, E64, O15.

1 Introduction

Poverty predictions are used in a wide variety of contexts where information on income, consumption or expenditure is scarce. Examples include proxy-means tests for the purpose of targeting (Coady et al. (2004), Brown et al. (2018)), poverty mapping for the purpose of estimating poverty for small geographical areas (Elbers et al. (2003)) and cross-survey imputations for estimating poverty for periods of time or populations characterised by scarce data on income (Dang et al. (2014), Douidich et al. (2016)). In all these cases, the base model of choice for estimating poverty is an OLS model.

In principle, predicting poverty can be done with either continuous or categorical dependent variable models such as OLS or Logit models. More recently, economics has started to experiment with machine learning models such as random forest or LASSO models as viable alternatives to address a variety of prediction problems (Varian (2014), Mullainathan and Spiess (2017), Athey and Imbens (2019)). A recent 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 poverty estimations (Fitzpatrick et al. (2018)).[1]

This paper compares the performance of basic econometric models, such as OLS and Logit models, with that of basic machine learning models, such as Random Forest and LASSO models, in the context of poverty predictions. It shows that it is unwise to express a preference for any model before this comparison is made in the context of a specific data set. Before opting for any particular prediction model, it is important to compare the performance of alternative models using different objective functions and possibly perform a stochastic dominance analysis across coverage curves.

2 Models

For simplicity, we use the two most common econometric models (OLS and Logit) and the two most popular machine learning models used by economists (Random Forest and LASSO) with each of the two machine learning models used with a continuous and a dichotomous dependent variable. We are comparing therefore a total of six models, three with a continuous dependent variable (income) and three with a dichotomous dependent variable (poor/non-poor) focusing on out-of-sample predic-

[1] See details of this competition on GitHub: https://github.com/worldbank/ML-classification-algorithms-poverty.

tions.[2] Predicting household poverty with these two classes of models requires three steps defined as 'Modelling', 'Prediction' and 'Classification'. In the case of OLS and Logit models the three steps are described as follows:

Step 1 - Modelling: $W_i = \alpha + \beta_1 X_i + \eta_i + \epsilon_i$ and $P_i = \delta + \gamma_1 X_i + \nu_i + \psi_i$, 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 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.

Step 2 - Prediction: $\widehat{W_i} = \widehat{\beta_1} X_i$ and $\widehat{P_i} = \widehat{\gamma_1} X_i$, where $\widehat{W_i}$, $\widehat{P_i}$ are predicted welfare or poverty.

The third and final step is to divide the population into estimated poor and non-poor groups. For this purpose, the two models critically differ. Under the OLS model, the poverty line is used after the second step to separate the estimated poor from the estimated non-poor. Under the Logit model, the same poverty line is used to separate the true poor from the true non poor to construct the poor dichotomous variable in step 1, whereas an arbitrary probability cutpoint is used to separate the estimated poor from the estimated non poor after step 2. Therefore, Step 3 can be described as follows:

Step 3 - Classification: $if\ \widehat{W_i} < z : i = poor; else : \ i = nonpoor$ and $if\ \widehat{P_i} > prob* : i = poor; else : \ i = nonpoor$, where z is the poverty line with $W_{min} \leq z \leq W_{max}$ and $prob*$ is an arbitrary probability cutpoint with $0 \leq prob* \leq 1$.

All poverty prediction models (econometric or machine learning) 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 (Table1). The matrix divides the population into four groups and can be used to construct targeting ratios that will be instrumental to define the objective function to optimize.

3 Objective Functions

The primary objective of a poverty reduction program is to minimize poverty. If, for the sake of simplicity, we consider the poverty rate as our primary objective to minimize, the corresponding value to minimize in the confusion matrix would be the False Negatives (FN). More generally, economists and computer scientists aim at minimizing Type I (False Positives or leakage) and Type II (False Negatives or

[2]The out-of-sample predictions are obtained by splitting the data set in two equally sized randomly selected sub-sets of data. The prediction coefficients are estimated from one data sub-set and used to make predictions in the other. We also run in-sample predictions with similar results in terms of heterogeneity of outcomes.

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.

undercoverage) errors. Since a two objective function is complex to optimize, most scholars reduce the problem to one objective by prioritizing one of the two errors. Welfare economists tend to prioritize the coverage rate (the reciprocal of Type II error) and pick the model that maximize coverage when Type I error (leakage) is kept fixed (Verme and Gigliarano (2019)). This paper follows this strategy.

The simplest approach with a single-objective function is to minimise a loss function such as the Mean Squared Error ($MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$), the Mean Absolute Error ($MAE = \frac{1}{N}\sum_{j=1}^{n}|y_j - \hat{y}_j|$) or combinations of the two such as the Huber loss function. These functions are routinely used by welfare economists for selecting prediction models but they are suitable for comparing different flavors of a single model such as the OLS model whereas they are not suitable for comparing different models such as an OLS and a Logit model. If we wish to compare different models, the most practical approach is to compare objective functions derived from the confusion matrix such as coverage, leakage, specification, precision, accuracy, Chi-squared or F2 ratios. For illustrative purposes and keeping in mind that the coverage rate is our primary objective function, we can use all these functions to compare the performance of the six models proposed.

It is also important to compare the sensitivity of models' performance when the poverty line changes. To compare poverty prediction models, one has first to set the poverty line. Each poverty line corresponds to a single confusion matrix resulting in a specific optimum for each of the functions considered allowing for models' comparison and choice of the optimal model. However, this choice of model may not be consistent if the poverty line changes. Therefore, when comparing models, one may want to conduct a stochastic dominance analysis by varying the poverty line. In the specific case of targeting the poor where the primary objective is to maximise coverage (leakage being equal), the proper curve for stochastic dominance analysis is the coverage curve plotted on the coverage-leakage plan.

4 Empirical Application

We take a real dataset of a middle income country and strip it of unnecessary variables and problematic observations to create a dummy dataset.[3] The final data set contains 7,062 observations and seven variables (income and six independent variables: gender, age, marital status and skills of the head of the household, household size and urban-rural location). The initial poverty line is set conveniently at the median value.[4] All models are run with the most basic specifiction allowed by the software. Here we are not seeking to optimize each model, simply to compare models across their simplest specification. Performance scores are therefore sub-optimal but our focus is on the *relative* performance across models. [5]

Table 2 compares the models.[6] The top of the table reports the predicted poverty rate and the *t-tests* for means difference between the true and predicted poverty rates and the bottom of the table reports all the objective functions used for models' comparison. The table shows that the best performing model in predicting poverty is the Logit model. This model is also the best performing model in terms of coverage together with the dichotomous LASSO model whereas the best performing model in terms of leakage is the LASSO continuous model. The other objective functions are also not consistent in ranking the models, although the logit model performs better than other models on most counts.

Figures 1 and 2 provide the stochastic dominance analysis by comparing coverage curves depicting trade-offs between coverage and leakage rates as we vary the choice

[3]As the data cleaning results in poverty rates that are not representative of the country selected, we will keep the country anonymous.

[4]When predicting poverty, a poverty line that is too low or too high will give too much importance to the poor or the non-poor. As the objective of the paper is to compare models, the median value avoids this problem. The sensitivity analysis that follows addresses the issue of comparing models when the poverty line changes.

[5]All estimations are conducted in STATA. In addition to the *reg* and *logit* commands for the OLS and Logit models, we use the *randomforest* comand for Random Forest and the *lasso* command for the LASSO model. The models' syntax in STATA is the following: 1) reg inc male age marstat skills urban hhsize; 2) randomforest inc male age marstat skills urban hhsize, type(reg) iter(100); 3) lasso linear inc male age marstat skills urban hhsize, selection(plugin); 4) logit truepoor male age marstat skills urban hhsize; 5) randomforest truepoor male age marstat skills urban hhsize, type(class) iter(100); 6) lasso logit truepoor male age marstat skills, selection(plugin).

[6]The objective functions are defined as follows: TPR=CR=Coverage Rate=Sensitivity Rate=TP/(TP+FN); TNR=Specificity Rate=TN/(TN+FP); FPR=Inclusion Rate=Leakage Rate=FP/(FP+TN); FNR=Exclusion Rate=Undercoverage Rate=FN/(FN+TP); chi2=Chi-squared; chi2lr=Chi-squared likelihood ratio; Precision=TP/(TP+FP); Accuracy=(TP+TN)/(TP+FP+TN+FN); F2=5*TP/(5*TP+4*FN+FP). All ratios are in percentage terms.

Table 2: Models' Comparison

	Continuous Dependent Var.			Dichotomous Dependent Var.		
	OLS	Ran.For.	LASSO	Logit	Ran.For.	LASSO
TruePov	48.16	48.16	48.16	48.16	48.16	48.16
PredPov	41.15	49.66	41.1	49.34	50.11	49.45
Diff(tstat)	7.41	-1.51	7.48	-1.26	-1.93	-1.38
TN	1377	1189	1378	1257	1149	1253
FP	428	616	427	548	656	552
FN	672	564	673	507	588	507
TP	1005	1113	1004	1170	1089	1170
TPR=CR=Sens.	59.93	66.37	59.87	69.77	64.94	69.77
TNR=Spec.	76.29	65.87	76.34	69.64	63.66	69.42
FPR=IE=LR	23.71	34.13	23.66	30.36	36.34	30.58
FNR=ER=UR	40.07	33.63	40.13	30.23	35.06	30.23
chi2	470.83	361.48	470.92	540.1	284.31	534.01
chi2lr	481.95	368	482.05	554.94	288.34	548.53
Precision	70.13	64.37	70.16	68.1	62.41	67.94
Accuracy	68.41	66.11	68.41	69.7	64.27	69.59
F2	61.72	65.96	61.68	69.43	64.42	69.4

of poverty line (expressed in deciles of income in the the figure). This is similar to a stochastic dominance approach with the difference that we are not comparing CDFs of income but coverage curves.[7]

Figure 1 compares the coverage curves across the three continuous dependent variables models. It shows that there is no absolute dominance of one model over the others as the curves cross each other repeatidly for low and high poverty lines. Similalry, Figure 2 compares the coverage curves across the dichotomous dependent variable models and finds the curves to cross each other with very low or very high poverty lines. In this last case, we can see that the Logit and LASSO models dominate the Random Forest model for most poverty lines but they do cross each other along most of the distribution with no clear dominance of one of the two models over the other. Finally, if we compare the best performing continuous and dichotomous dependent variable models (OLS and Logit models respectively), we find that the two curves intersect (not shown in figures).

Similalry to ROC curves, the Area Under the Curve (AUC) can be used as the function to maximize when comparing models. It is evident that two curves may have identical AUC but different shapes if they intersect multiple times. A stochastic dominance analysis provides more granular information than AUC values by showing whether the curves intersect and where.

5 Conclusions

The paper has compared basic econometric and machine learning models used by economists to predict poverty at the household level. It showed that no model can be said to be superior to others *ex-ante*, before models are tested in the context of a specific income distribution. *A priori*, we cannot predict which model outperforms other models in the context of a country-specific targeting exercise. The choice of the optimal model depends on the location of the poverty line, the choice of objective function and the particular income distribution at hand. Unlike current practices, it is essential to test alternative models and perform stochastic dominance analysis before selecting the optimal model to use for delivering assistance to the poor.

[7]The coverage curve is the same as the ROC curve. It has been shown that there is an equivalence between ROC curves and the CDFs of probability measures on the unit interval (Gneiting and Vogel (2018)).

Figure 1: Models' Comparison (Continous Dep. Var.)

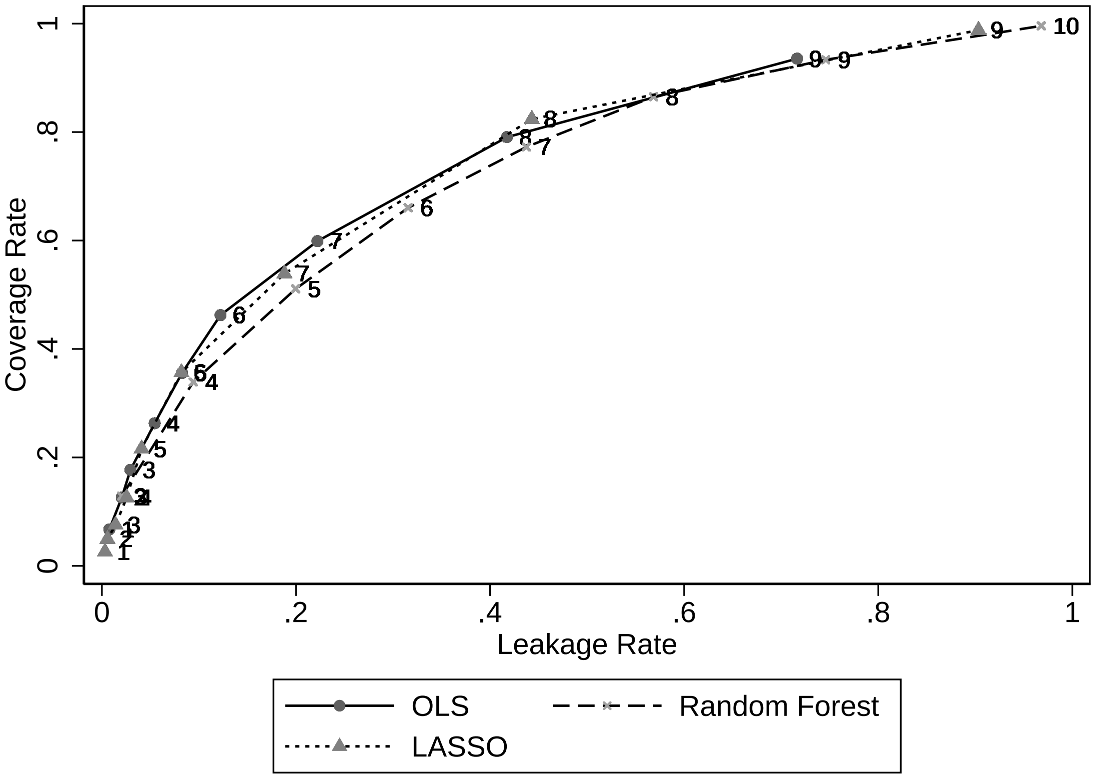
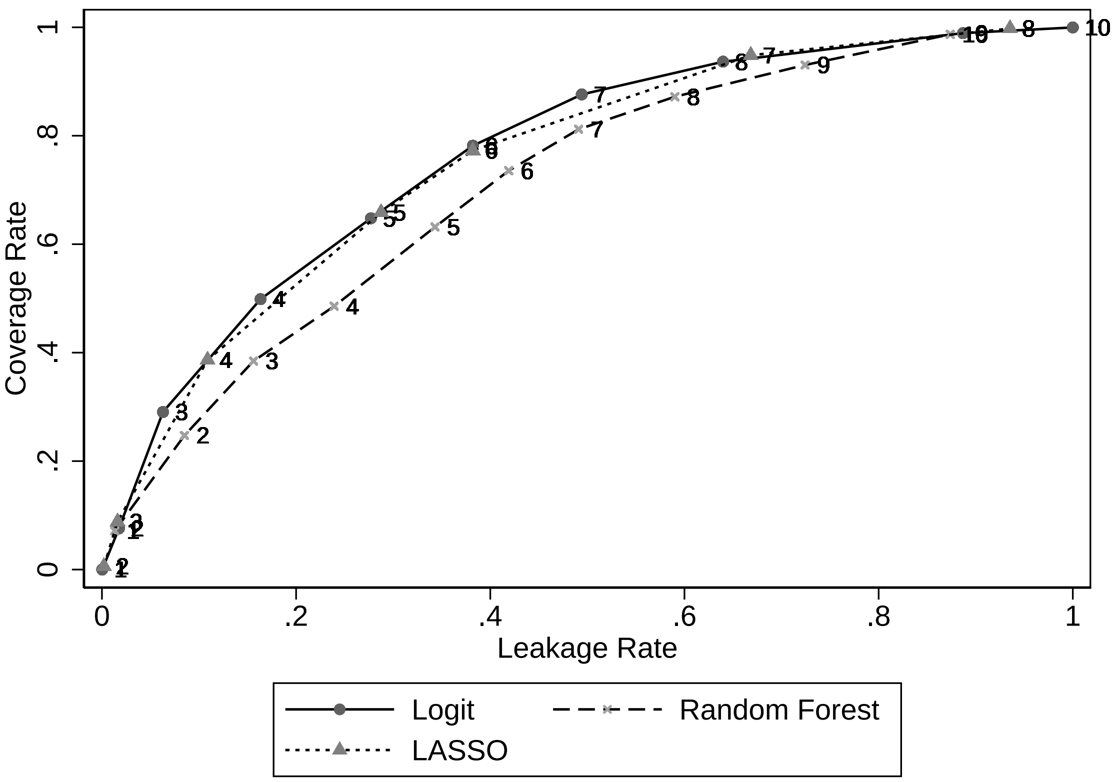


Figure 2: Models' Comparison (Dichotomous Dep. Var.)



References

Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. *arXiv* (1903.10075).

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.

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

Dang, H.-A., P. Lanjouw, J. Luoto, and D. McKenzie (2014). Using repeated cross-sections to explore movements into and out of poverty. *Journal of Development Economics 107*(C), 112–128.

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.

Fitzpatrick, C. A., P. Bull, and O. Dupriez (2018). Machine learning for poverty prediction: A comparative assessment of classification algorithms. *Wired at: www.github.com*.

Gneiting, T. and P. Vogel (2018). Receiver operating characteristic (roc) curves. *arXiv* (1809.04808).

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31* (2), 87–106.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28* (2), 3–28.

Verme, P. and C. Gigliarano (2019). Optimal targeting under budget constraints in a humanitarian context. *World Development 119* (C), 224–233.