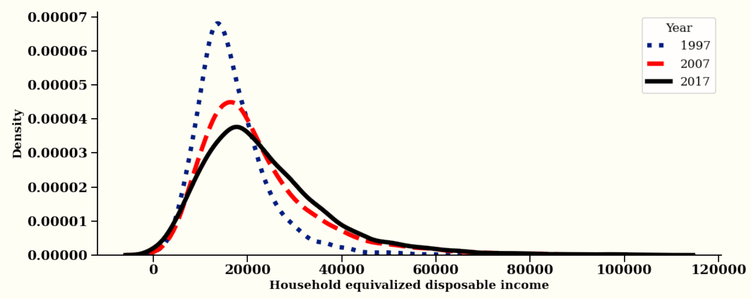
PREDICTING VULNERABILITY TO POVERTY WITH MACHINE LEARNING

Alemayehu TAYE, Prof. Conchita D'AMBROSIO, and Prof. Alexandre TKATCHENKO
University of Luxembourg
May 19, 2021

Motivation

- Effective public policy that aims to reduce poverty should consider the household vulnerability to poverty, the ex-ante risk of falling into poverty in the future [1].
- Between one year and the next, many people move into or out of poverty, hence measures of who is poor now are imperfect guides to who will be poor next period.
- Although modeling socioeconomic data is very complex, the vulnerability to poverty literature has not yet taken advantage of machine learning (ML).
- Therefore in this project, we aim to address if a fully data-driven predictive modeling help us to accurately target the disadvantaged groups.

Figure 1: *Evolution of HH income distribution in Germany*



Dataset

- The German socio-Economic Panel (**SOEP**) a representative longitudinal data of private households in Germany since 1984.

Empirical strategy

- Before determining **households vulnerability to poverty** (V_{ht}), we model the inter-temporal and cross-sectional determinants of household disposable income as follows:

$$y_{h,t} = f(y_{h,t-1}, \boldsymbol{X}_h, \alpha_h, \epsilon_{ht}) \tag{1}$$

where $y_{h,t-1}$ is lag of the real household equivalized income and $\boldsymbol{X}_h$ the vector of observable household features.

- In the intermediate stage we identify households' current vulnerability status to future poverty as follows:

$$\hat{V}_{h,t} = \begin{cases} 1 & \text{if} \quad \tilde{y}_{h,t+1} = \hat{f}(y_{ht}, \boldsymbol{X}_h, \alpha_h) \leq \tilde{z} \\ 0 & \text{if} \quad \tilde{y}_{h,t+1} = \hat{f}(y_{ht}, \boldsymbol{X}_h, \alpha_h) > \tilde{z} \end{cases} \tag{2}$$

where $\tilde{y}_{h,t+1}$ is the **forecast** of household equivalized income in period $t+1$ and $\tilde{z} = 0.6 * \text{median}(\tilde{y}_{h,t+1})$.

- ***Sensitivity***: the probability of detecting vulnerable households in period t who actually become poor in period $t+1$.
- ***Accuracy***: the proportion of households in the sample N that were correctly classified as vulnerable ($\hat{v}_h = 1$) and non-vulnerable ($\hat{v}_h = 0$).

Models

1. **Linear Regression** (baseline model).
2. **LASSO**: linear model with $\boldsymbol{L}_1$ regularization
3. **Ridge Regression**: linear model with $\boldsymbol{L}_2$ regularization
4. **Random Forest** [2]
5. **Gradient Boosting Trees** [3]
6. **Neural Network** (Multi-layer Perceptron)[4]

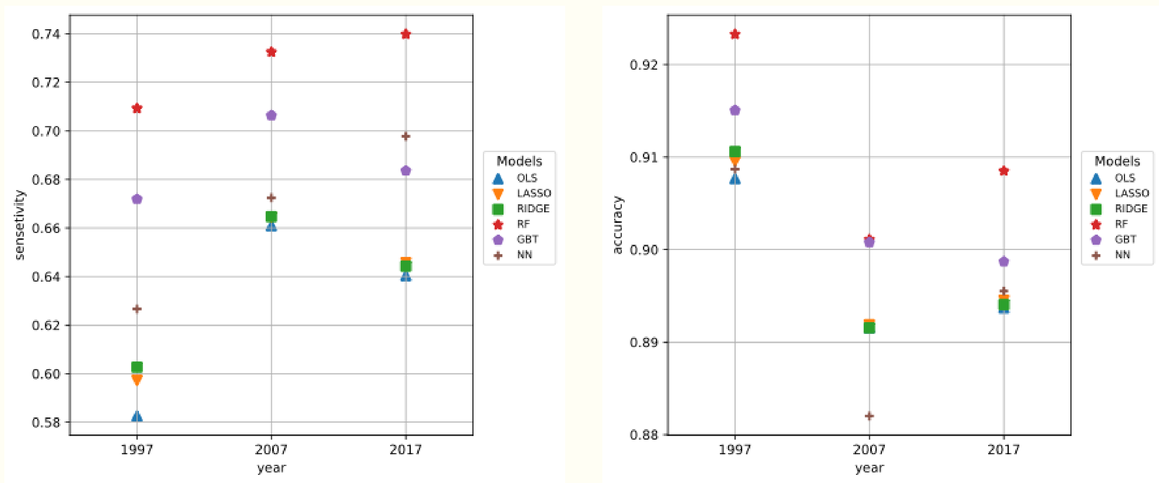
- Each algorithms has been grid search with 10-fold CV on the training set. We report the models generalizability on the holdout test set.

Results

Tabel 1: *Summary of ML algorithms performance in modeling household equivalized income*

Datasets	Model	R-Square		RMSE		Hyperparameters
		Trainset	Testset	Trainset	Testset	
1997	OLS	0.717	0.722	4966.35	4974.92	
	Lasso	0.716	0.723	4966.50	4972.50	α = 0.447
	Ridge	0.717	0.721	4966.35	4974.91	α = 0.016
	Random Forest	0.890	0.724	3091.98	4948.75	max_depth = 10, max_features = 28
	GBT	0.761	0.728	4560.34	4912.99	n_estimator =50, max_depth = 3, learning_rate = 0.119
	Neural Network	0.763	0.712	4549.28	5063.07	n_neuron = {200(layer 1), 200(layer 2)}, α = 0.006
2007	OLS	0.789	0.816	6041.60	5532.41	
	Lasso	0.788	0.816	6050.65	5530.84	α = 0.711
	Ridge	0.789	0.815	6046.51	5542.41	α = 0.006
	Random Forest	0.904	0.812	4072.61	5589.38	max_depth = 10, max_features = 30
	GBT	0.823	0.815	5527.66	5542.16	n_estimator =250, max_depth = 3, learning_rate = 0.042
	Neural Network	0.808	0.805	5762.77	5695.45	n_neuron = {200(layer 1), 50(layer 2)}, α = 0.007
2017	OLS	0.786	0.760	5655.218	5842.15	
	Lasso	0.786	0.761	5655.303	5841.05	α = 0.232
	Ridge	0.786	0.760	5655.218	5842.15	α = 0.008
	Random Forest	0.914	0.766	3576.727	5774.89	max_depth = 11, max_features = 29
	GBT	0.826	0.767	5096.524	5757.74	n_estimator =200, max_depth = 3, learning_rate = 0.057
	Neural Network	0.814	0.745	5279.747	6029.81	n_neuron = {200(layer 1), 120(layer 2)}, α = 0.004

Figure 2: *Performance of models in estimating households vulnerability to poverty*



Experiments /Robustness check

1. We experiment with two different set of features – i) pre-engineered 30 features from[5] and ii) Our own construct of large set of features (70).
2. We check if the predictions are sensitive to the different survey years (1997, 2007, and 2017).
3. We experiment two different vulnerability cutoff points, i) 0.6*median(y_{ht}) and ii) 0.6*median($\tilde{y}_{h,t+1}$).

Conclusion

Optimized tree-based algorithms, RF in particular, show high potential in predicting vulnerability to poverty:

- The positive gain holds for all three different survey years.
- The result holds with both sets of features. However, we obtained more pronounced gain with our set of features.
- The result is robust to the two alternative vulnerability cutoffs , but we suggest the endogenous vulnerability cutoff.

The **powerful predictors** of households disposable income include:

- Previous year household equivalized income.
- Characteristics of the household head such as, education status, hours worked, marital status, occupation class , gender, health status.
- Households composition (by age and by activity status in the labor market).
- Location of the household (region of residence)

Acknowledgement

The Doctoral Training Unit **Data-driven computational modelling and applications** (DRIVEN) is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781). https://driven.uni.lu

References

[1] Malcolm Gillis, Carl Shoup, and Gerardo P Sicat. *World development report 2000/2001-attacking poverty*. The World Bank, 2001.

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] Jerome H Friedman. "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[4] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

[5] Maike Hohberg et al. "Vulnerability to poverty revisited: Flexible modeling and better predictive performance". In: *The Journal of Economic Inequality* 16.3 (2018), pp. 439–454.