Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Predicting transitions into and out of poverty using machine learning

by Joep Burger and Jan van der Laan

Release date: October 15, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Predicting transitions into and out of poverty using machine learning

Joep Burger, and Jan van der Laan[1]

Abstract

The increasing size and richness of digital data allow for modeling more complex relationships and interactions, which is the strongpoint of machine learning. Here we applied gradient boosting to the Dutch system of social statistical datasets to estimate transition probabilities into and out of poverty. Individual estimates are reasonable, but the main advantages of the approach in combination with SHAP and global surrogate models are the simultaneous ranking of hundreds of features by their importance, detailed insight into their relationship with the transition probabilities, and the data-driven identification of subpopulations with relatively high and low transition probabilities. In addition, we decompose the difference in feature importance between general and subpopulation into a frequency and a feature effect. We caution for misinterpretation and discuss future directions.

Key Words: Classification; Explainability; Gradient boosting; Life event; Risk factors; SHAP decomposition.

1. Introduction

In 2015 the United Nations set 17 Sustainable Development Goals for 2030, one of which was to end poverty in all its forms everywhere. More locally, Dutch municipalities are interested in improving policies aimed at reducing and preventing poverty. To reach these goals it is important to know risk factors that drive transitions into and out of poverty. Several are already known, such as educational attainment, household composition, unemployment rate and access to public services (e.g. Kemp et al., 2004). Typically, poverty research is done in an econometrically framework (Andriopoulou and Tsakloglou, 2011). The ever increasing size and richness of digital data allow for modeling more complex, nonlinear relationships and interactions, for which machine learning seems more suitable (Breiman, 2001). Our research question therefore was: can we obtain new insights on risk factors when we apply machine learning to a rich set of register data? This research was requested by the Dutch Ministry of the Interior.

2. Data

The target population consisted of all adults in private households residing in the Netherlands on the 31st of December of a given year for which the household income and assets were known in both that year and the next, thus excluding minors, institutional households, student households and recent migrants. The population was split into two: the poor population consisting of about 0.6 million people and the non-poor population consisting of about 12 million people. The target variable was an indicator for poverty in the next year. In this way we could model the transitions into and out of poverty. By our definition, a person was considered poor in a given year if the standardized disposable household income was below the poverty threshold AND the standardized disposable household assets were below half the poverty threshold, which is about the maximum amount to qualify for social security (van den Brakel and Otten, 2019). We combined the populations of two reference years (2013 and 2017), doubling the datasets and increasing generalizability over time.

We used the Dutch system of social statistical datasets, which is a system of interlinked and standardized registers and surveys (Bakker et al., 2014). From this, we derived over 500 features about persons and households. Regional features

[1]Joep Burger, Statistics Netherlands, CBS-weg 11, PO Box 4481, 6401 CZ Heerlen, the Netherlands (j.burger@cbs.nl); Jan van der Laan, Statistics Netherlands, Henri Faasdreef 312, PO Box 24500, 2490 HA The Hague, the Netherlands (dj.vanderlaan@cbs.nl)

were excluded because the machine learning algorithm could not handle well the hierarchical structure in the data. We derived features on income and expenditure. The feature set included demographic, socioeconomic, health, child welfare and crime variables. Not only the present status but also changes in the past three years (life events) were considered.

3. Method

The labeled dataset was used to learn the relationship between the features and the label. In case of the poor population this label was the indicator for staying poor, which happens on average to over 70% of the poor population. In case of the non-poor population, the label was the indicator for becoming poor, which happens on average to about 1% of the non-poor population. To learn this relationship, we applied gradient boosting using the implementation by Chen and Guestrin (2016). Gradient boosting is a popular decision tree-like machine learning algorithm, where subsequent trees are trained on residuals of the predictions by earlier trees. Compared with for instance logistic regression it scales easier to many features and is more flexible in learning nonlinear relationships and complex interactions. Nested cross validation was applied to tune hyperparameters and test model performance.

We picked three quality metrics that capture different aspects of the model: the area under the Receiver Operating Characteristic curve (AUC), Matthews Correlation Coefficient (MCC) and the harmonic mean of recall and precision of the positive class (F_1^+). A person was classified as poor if the predicted probability exceeded a cutoff c, which was optimized per metric. To compare model performance between models trained on data that differ in class imbalance, min-max normalization (mmn) was applied using as minimum the score obtained when guessing the positive class ('stays/ becomes poor') with a probability equal to the observed fraction that stays/ becomes poor in the training set (Burger and Meertens, 2020).

Although machine learning is more flexible in learning the output labels than regression models, it is also focused on prediction quality rather than explainability. To better understand how the models come to their predictions, we applied two methods: SHAP and global surrogate models (see Molnar, 2019). A SHAP value is a weighted sum of marginal contributions of a feature to the predicted value of a person. It provides insight into the learned relationship between features and predictions. A global surrogate model is a simpler model fit to the predictions of the complex model. It provides insight into the most important features and allowed us to define data-driven subpopulations for more in-depth study.

4. Results

4.1 Prediction quality

On the test sets, the model for the poor population scores lower on AUC than the model for the non-poor population, but it scores better on MCC and positive F_1 (Table 4.1). More importantly, all normalized scores are way above zero, which means that the models outperform random guessing and have learned from the input features.

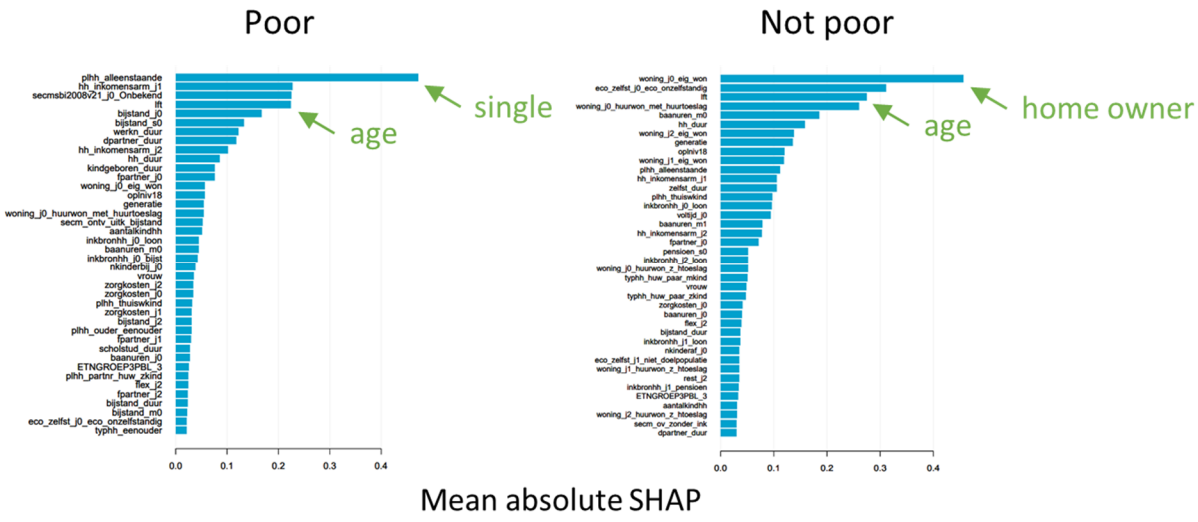
Table 4.1
Model performance

Metric	Population	
	Poor	Not poor
AUC^{mmn}	0.67	0.90
MCC(c*)	0.48	0.42
$F_1^{+,mmn}$	0.53	0.42

4.2 Feature importance

If we average the absolute SHAP value over persons we can rank features by importance. According to this method, the most important feature for the poor population is the indicator for being single and for the non-poor population the indicator for home ownership (Fig. 4.2). Age is an important feature in both models. Most of the 500+ features have little effect on the model predictions. A description of the feature labels can be found in van der Laan et al. (2021).

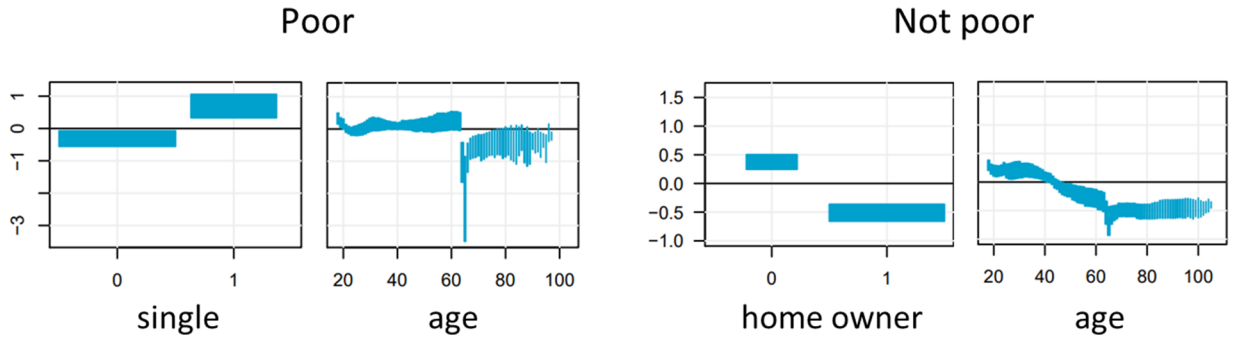
Figure 4.2
Feature importance



4.3 SHAP

In Figure 4.3 we zoom in on the distribution of the SHAP values by feature. The height of the bars cover 80% of the SHAP values, the width of the bars represent the relative frequency of the feature value in the population. Now we see that the indicator for being single is not only important but also that most people who are single have a higher probability of staying poor than people who are not single. Similarly, we see that most home owners have a lower probability of becoming poor than people in rental houses. In both models, reaching the retiring age dramatically lowers the probability of staying or becoming poor. Elderly who do not qualify for a pension still have a lower than average poverty risk. Finally, the poverty risk stays fairly constant with age in the poor population but drops considerably with age in the non-poor population. In summary, the SHAP values provide detailed insight in both the nature and strength of the non-linear relationship between feature and estimated poverty risk.

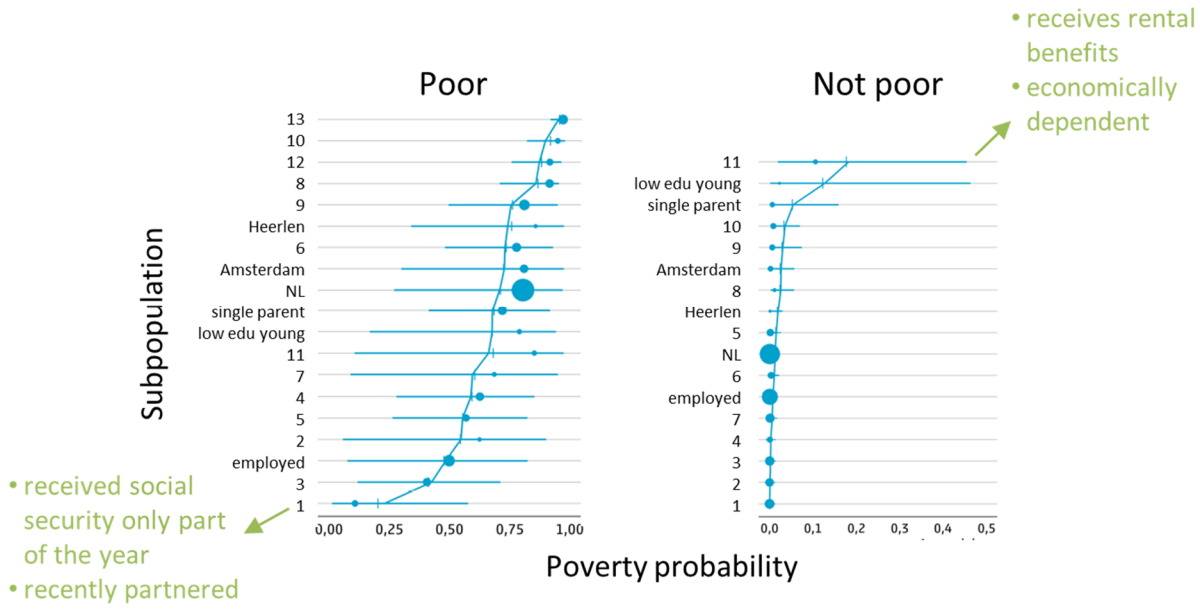
Figure 4.3
SHAP values for two important features



4.4 Global surrogate models

In Figure 4.4 you see the observed and predicted poverty probabilities for the subpopulations from the global surrogate models. The irregular blue line connects the observed fractions, the vertical dashes very close by are the average predicted probabilities, the circles are the median predicted probabilities with their area proportional to the size of the subpopulation, and the horizontal blue lines indicate the 10th and 90th percentiles of the predicted probabilities. This data-driven approach has resulted in subpopulations that are more distinct than the expert-driven approach. For instance, poor people who do not receive social security or only part of the year and have recently partnered (poor subpopulation 1) have on average lower poverty risks than for instance poor people who are employed (an expert-defined subpopulation). Similarly, non-poor people who receive rental benefits and are not economically independent (non-poor subpopulation 11) have on average higher poverty risks than for instance non-poor, lowly educated young people (another expert-defined subpopulation). Exact definitions of all subpopulations can be found in van der Laan et al. (2021).

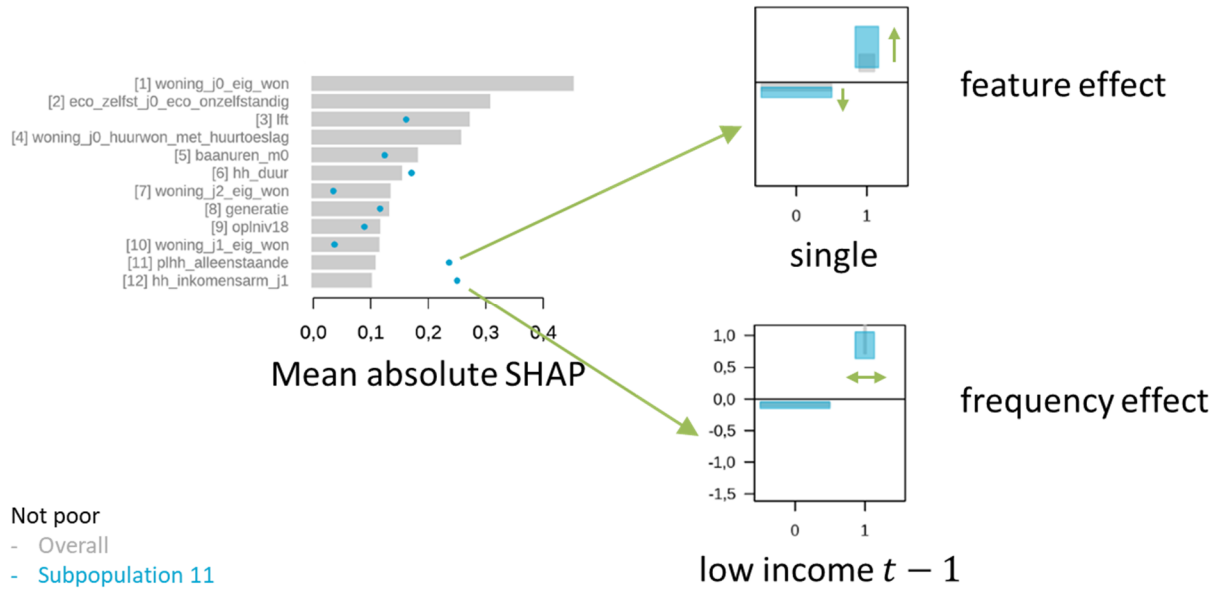
Figure 4.4
Global surrogate models



4.5 SHAP decomposition

In Figure 4.5 we compare the feature importance between the general population (non-poor population; gray bars) and a subpopulation (non-poor subpopulation 11; blue dots). Two features are clearly more important in this subpopulation: the indicator for being single and the indicator for having a low income in the previous year. The indicator for being single is more important in the subpopulation mostly because its effect is stronger. The indicator for having a low income in the previous year is more important in the subpopulation mostly because the subpopulation contains more people who had a low income in the previous year. In general, the difference in SHAP value between subpopulation and general population can be decomposed into a feature effect and a frequency effect. The derivation can be found in van der Laan et al. (2021).

Figure 4.5
SHAP decomposition



5. Discussion

We would like to highlight three points for discussion. First, this approach is based on observational data, so policy makers have to realize that an important feature is not necessarily a causal effect. For instance, receiving rental benefits might be a good predictor for becoming poor but is unlikely to cause poverty. Lowering or abolishing rental benefits will not be an effective poverty prevention measure. Second, we believe that SHAP is a useful method to rank features, but other methods might result in different ranking of features. Third, this approach is suitable to identify risk factors but not to profile individuals.

We conclude that the model predictions are reasonably good. Applying machine learning to register data allows for ranking potential risk factors and mapping detailed nonlinear relationships. The SHAP decomposition of data-driven subpopulations and the inclusion of life events in the recent past provide additional insight for policy makers. However, with the predictive power of machine learning comes great responsibility in terms of explainability.

Future research could focus on incorporating the dependency in hierarchical data, maybe by developing a machine learning equivalent of multilevel regression models. Policy makers are probably interested in causal models, model fairness and actionable features, which have not been addressed here. Counterfactuals is another way of making machine learning models more transparent. Finally, a similar approach could be applied to model long-term poverty or well-being.

References

Andriopoulou, E., and P. Tsakloglou (2011), "The determinants of poverty transitions in Europe and the role of duration dependence", IZA Discussion Paper No. 5692, https://ssrn.com/abstract=1842089.

Bakker, B. F. M., J. van Rooijen, and L. van Toora (2014), "The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics", *Statistical Journal of the IAOS*, 30, pp. 411–424.

Brakel, M. van den, and F. Otten (2019), "Poverty and social exclusion", report (in Dutch), The Hague, the Netherlands: Statistics Netherlands. https://www.cbs.nl/-/media/_pdf/2019/50/armoede-en-sociale-uitsluiting-2019.pdf.

Breiman, L. (2001), "Statistical modeling: The two cultures", *Statistical Science*, 16(3), pp. 199–231.

Burger, J., and Q. Meertens (2020), "The algorithm versus the chimps: On the minima of classifier performance metrics", *Proceedings of the BNAIC/BeneLearn2020*, pp. 38–55.

Chen, T., and C. Guestrin (2016), "XGBoost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Kemp, P. A., J. Bradshaw, P. Dornan, N. Finch, and E. Mayhew (2004), "Routes out of poverty: A research review", research report, York: Joseph Rowntree Foundation, https://eprints.whiterose.ac.uk/73260/.

Laan, J. van der, J. Burger, M. Detiger, N. Schalken, W. van Andel, M. van den Brakel, and S. Tan (2021), "Risk factors for transitions into and out of poverty", report (in Dutch), The Hague, the Netherlands: Statistics Netherlands, https://www.cbs.nl/-/media/innovatie/rapport-armoede.pdf.

Molnar, C. (2019), *Interpretable Machine Learning*, https://christophm.github.io/interpretable-ml-book/.