Predicting Poverty Susceptibility Using U.S. Census Microdata

Authors: Marianne Sawires, Ingrid Altamirano, Vanessa Scott

Poverty is a critical public health issue with far-reaching implications for physical health, mental health, housing stability, and access to social services. Our team seeks to utilize the open-source United States Census data from the American Community Survey (ACS) which provides a rich source of information on household finances, demographic characteristics, and participation in government-funded assistance programs. Leveraging these data through data science methods offers an opportunity to better understand the structural and individual-level factors associated with economic vulnerability. Our interdisciplinary team, with backgrounds in psychiatry, social work, finance, and data science, brings complementary perspectives to examining poverty within a public health framework.

The primary data science problem addressed in this project is to identify and quantify the factors that contribute most strongly to an individual's or household's likelihood of experiencing poverty. By using historical multi-year data, our model aims to accurately predict poverty susceptibility for the 2024 survey year. This approach treats the 2024 data as an "out-of-time" test set to evaluate how well the learned socio-economic risk factors generalize across changing economic landscapes.

For this supervised learning task, the input features consist of demographic characteristics, household composition, employment indicators, income measures, and participation in government assistance programs. The primary outcome variable will be the Census-defined income-to-poverty ratio variable, POVPIP, which captures economic status relative to the federal poverty threshold which can be used for classification, enabling standard evaluation. We will leverage feature importance tasks as a method for selecting from the 280 features available.

Machine Learning fits this problem because poverty is driven by the numerous factors listed above. These relationships are often nonlinear and high dimensional, so an ML model is perfect for learning these complex patterns and capturing which features are most predictive, while linear or rule-based approaches can miss these complexities. We will be using American Community Survey 1-Year Estimates Public Use Persons Microdata, provided by the U.S Census Bureau. It contains at least 286 variables and over 2 million records for the past 5 years. This dataset will provide enough records to create separate training, validation and test splits, and even multiple test sets if needed.

A lot of the existing research has approached the problem of who is at risk of being in poverty by using longitudinal survey data such as Panel Study of Income Dynamics (PSID), the Survey of Income and Program Participation (SIPP), and the Health and Retirement Study (HRS), applying econometric models to study transition in and out of poverty (McKernan & Ratcliffe, 2005; Clark, Lusardi, & Mitchell, 2024). These studies highlight the role of factors such as employment changes, household composition, and health shocks in driving poverty transitions. Recent work has shown that ML methods, such as tree-based models, can improve predictive performance (Burger & van der Laan, 2021; Verme, 2024). In our project, we will likely be testing similar models, focusing on prediction and feature importance instead of casual inference.

Moving onto the model build process, we will start with defining a baseline model and implement a regularized multivariate logistic regression model to predict poverty status using the Census-defined income-to-poverty ratio (POVPIP). Input features will be drawn from a broad set of demographic, household, employment, housing, disability, and public assistance variables available in the ACS microdata, with feature importance analyses used to identify which characteristics are most strongly associated with poverty status. Regularization and class weighting will be used to address multicollinearity and class imbalance.

Some of the models we are particularly interested in exploring are the Random Forest classifier, and XGBoost given the large number of variables and the complexity of the underlying relationships in the data. Poverty is influenced by many interacting socioeconomic factors, and Random Forests and XGBoost are well suited to this setting because they can capture complex, non-linear relationships while reducing the impact of noisy or less informative features through ensemble averaging. This makes them a strong choice for high-dimensional census microdata. We will compare performance against a multivariate logistic regression baseline. If time permits, we are interested in using the FT-Transformer (Feature Tokenizer Transformer) model. By embedding each feature as a token and utilizing multi-head self-attention, the FT-Transformer can dynamically weigh global feature dependencies that traditional tree-based models may overlook. (Shavitt, Yoav, et al. 2021).

We intend to measure success by evaluating accuracy, recall, F1-score, and area under the precision-recall curve (AUC) on test data for predicting susceptibility to poverty. We will place particular emphasis on recall, as minimizing missed high-risk individuals is critical for identifying populations most vulnerable to poverty. Model robustness will be further evaluated through multiple train-test splits across 5 years of data, allowing us to assess how well learned socio-economic patterns generalize across changing conditions.

References

Burger, Ronelle, and Jacob van der Laan. 2021. *"Predicting Transitions in and out of Poverty using Machine Learning."* Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2021001/article/00003-eng.pdf?st=4uTd_yt_

Clark, Robert, Annamaria Lusardi, and Olivia S. Mitchell. 2024. "Financial Fragility during Economic Shocks." *Social Security Bulletin* 84, no. 3. https://www.ssa.gov/policy/docs/ssb/v84n3/v84n3p11.html.

McKernan, Signe-Mary, and Caroline Ratcliffe. 2005. "Events that Trigger Poverty Entries and Exits." *Social Service Review* 86 (supplement 1): https://econpapers.repec.org/article/blasocsci/v_3a86_3ay_3a2005_3ai_3as1_3ap_3a1146-1169.htm.

Shavitt, Yoav, et al. 2021. *"Revisiting Deep Learning Models for Tabular Data."* https://arxiv.org/pdf/2106.11959

Urban Institute. 2014. *Events that Trigger Poverty Entries and Exits.* Washington, DC: Urban Institute. https://www.urban.org/sites/default/files/publication/60726/410636-events-that-trigger-poverty-entries-and-exits.pdf.

Urban Institute. n.d. *Transitioning In and Out of Poverty.* Washington, DC: Urban Institute. https://www.urban.org/research/publication/transitioning-and-out-poverty.

Verme, Paolo. 2025. *"Predicting Poverty." arXiv* preprint, arXiv:2505.05958. https://arxiv.org/abs/2505.05958.