Household Consumption Imputation for Poverty Rate Prediction: A Comparison of Point Estimation and Distribution Modeling Approaches

Poverty Prediction Competition Analysis

January 2026

Abstract

This paper presents an analysis of methods for imputing household consumption and predicting poverty rates from survey data. We explore three main approaches: (1) point estimation using gradient boosting ensembles, (2) quantile regression for distribution modeling, and (3) mixture density networks (MDN) for full conditional distribution estimation. Using Leave-One-Survey-Out (LOSO) cross-validation to simulate the test scenario, we achieved CV scores of 9.14 with CatBoost ensembles, approximately 9.44 with quantile regression, and 7.72 with MDN. However, test performance revealed significant generalization gaps, with scores of 9.35, 10.02, and 12.04 respectively. We analyze the causes of this CV-test gap and discuss lessons learned for poverty prediction research.

1 Introduction

Accurate measurement of poverty is fundamental to effective development policy and resource allocation. The World Bank and national statistical agencies rely on household consumption surveys to estimate poverty rates, but such surveys are expensive and time-consuming to conduct. A common challenge in real-time poverty monitoring is that newer surveys may lack the detailed consumption modules needed to directly measure poverty, requiring imputation from available survey features.

This paper documents our approach to a poverty prediction competition that mimics this real-world challenge. The competition provides three training surveys with full consumption data and requires prediction on three test surveys where only survey features are available. The competition metric combines household-level consumption prediction error (10%) with weighted poverty rate prediction error (90%), emphasizing the importance of accurate aggregate statistics over individual predictions.

The key technical challenges include:

- **Distribution shift**: Test surveys may come from different time periods or populations than training surveys
- **Aggregation**: Poverty rates are aggregate statistics that depend on the entire predicted distribution, not just point predictions
- **Weighted inference**: Households have population weights that must be correctly handled
- **Limited validation**: Only 3 training surveys means LOSO CV has high variance

2 Related Work

2.1 Survey Imputation Methods

Imputation of missing consumption data in surveys has been studied extensively. Traditional methods include regression-based imputation and multiple imputation using chained equations. More recently, machine learning methods have been applied, including random forests and gradient boosting.

2.2 Gradient Boosting for Tabular Data

Gradient boosting decision trees (GBDT) remain state-of-the-art for tabular regression tasks. LightGBM, XGBoost, and CatBoost are widely used implementations with different optimization strategies. CatBoost is particularly effective for datasets with categorical features due to its ordered boosting and target encoding strategies.

2.3 Distribution Modeling

For poverty rate prediction, modeling the full conditional distribution of consumption is theoretically superior to point prediction. Quantile regression allows prediction at arbitrary quantiles, while mixture density networks (MDN) can model arbitrary conditional distributions using neural networks to predict parameters of mixture models.

3 Data Description

3.1 Dataset Overview

The dataset consists of six household surveys:

- **Training**: Surveys 100000, 200000, 300000 (approximately 35,000 households each)
- **Test**: Surveys 400000, 500000, 600000 (approximately 35,000 households each)

Each household has approximately 88 features after preprocessing, including:

- Demographic information (household size, age, gender of head)
- Education and employment indicators
- Housing characteristics (ownership, utilities, sanitation)
- Geographic indicators (urban/rural, region)
- Food consumption indicators (50 binary variables for items consumed in last 7 days)
- Survey sampling information (strata, weights)

3.2 Poverty Thresholds

The competition uses 19 poverty thresholds derived from the ventiles (5th to 95th percentiles) of the consumption distribution in Survey 300000:

Table 1: Poverty Thresholds (2017 USD PPP per capita per day)

\$3.17	\$3.94	\$4.60	\$5.26	\$5.88	\$6.47	\$7.06	\$7.70
\$8.40	\$9.13	\$9.87	\$10.70	\$11.62	\$12.69	\$14.03	\$15.64
		\$17.76			\$20.99	\$27.37	

3.3 Ground Truth Poverty Rates

The training surveys have the following poverty rates at selected thresholds:

Table 2: Ground Truth Poverty Rates by Survey

Threshold	Survey 100000	Survey 200000	Survey 300000
\$3.17 (5%)	6.74%	5.93%	4.98%
\$7.70 (40%)	41.98%	40.76%	39.99%
\$9.87 (55%)	57.44%	55.94%	55.01%
\$27.37 (95%)	95.42%	95.28%	95.00%

4 Methodology

4.1 Feature Engineering

We created 30+ engineered features in addition to the raw survey features:

1. **Household composition ratios**: children/household size, elderly ratio, adults ratio
2. **Food diversity score**: Sum of 50 food consumption indicators, representing dietary variety
3. **Infrastructure index**: Sum of water, toilet, sewer, electricity indicators
4. **Employment quality**: Formal worker ratio, workers per adult
5. **Log transforms**: Log of utility expenditure, log of weights (for skewed distributions)
6. **Household size features**: Squared household size, large household indicator
7. **Strata-based features**: High strata and low strata indicators
8. **Utility expense per person**: Normalized utility expenditure
9. **Region-urban interactions**: Cross-features of geographic variables

4.2 Point Estimation Models

We trained three gradient boosting models with the following configurations:

Table 3: Model Hyperparameters

Parameter	LightGBM	XGBoost	CatBoost
Learning rate	0.05	0.05	0.05
Max depth	-1 (unlimited)	7	7
Num leaves / Min samples	63	20	-
L2 regularization	0.1	0.1	3.0
Feature fraction	0.8	0.8	-
Estimators	1000	1000	1000

All models used log-transformed target ($\log(1 + \text{consumption})$) and MAPE loss function.

4.3 Quantile Regression

For distribution modeling, we trained quantile regression models at 19 quantiles corresponding to the poverty thresholds:

$$\tau \in \{0.05, 0.10, 0.15, \ldots, 0.95\}$$

CatBoost was used with quantile loss:

$$L_\tau(y, \hat{y}) = \sum_i \rho_\tau(y_i - \hat{y}_i)$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}_{u<0})$ is the check function.

Poverty rates were estimated by interpolating between predicted quantiles:

1. For each household i, predict quantile values $\hat{q}_i(\tau)$ for all τ
2. For threshold t, estimate $P(C_i < t)$ by finding where t falls among predicted quantiles
3. Aggregate using household weights: $\hat{R}(t) = \sum_i w_i \cdot \hat{P}(C_i < t) / \sum_i w_i$

4.4 Mixture Density Networks

We implemented a Mixture Density Network (MDN) that predicts parameters of a 5-component Gaussian mixture for each household:

$$p(c_i|x_i) = \sum_{k=1}^{5} \pi_{ik} \cdot \mathcal{N}(c_i; \mu_{ik}, \sigma_{ik}^2)$$

The network architecture:

- Input layer: 88 features (normalized)
- Hidden layers: $256 \to 256 \to 128$ (ReLU, BatchNorm, Dropout 0.2)
- Output heads: π (softmax), μ (softplus), σ (softplus + ϵ)

Training used negative log-likelihood loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\sum_{k=1}^{5} \pi_{ik} \cdot \mathcal{N}(c_i; \mu_{ik}, \sigma_{ik}^2) \right]$$

4.5 Competition Metric

The competition metric is a weighted combination:

$$\text{Score} = 0.9 \times \text{Rate\_Error} + 0.1 \times \text{Consumption\_Error}$$

where:

$$\text{Rate\_Error} = \frac{1}{S} \sum_s \frac{\sum_t w_t \cdot \left| \frac{\hat{r}_{st} - r_{st}}{r_{st}} \right|}{\sum_t w_t}$$

$$\text{Consumption\_Error} = \frac{1}{H} \sum_h \left| \frac{\hat{c}_h - c_h}{c_h} \right|$$

The threshold weights $w_t = 1 - |p_t - 0.4|$ prioritize accuracy near the 40% poverty rate.

5 Experiments and Results

5.1 Leave-One-Survey-Out Cross-Validation

We used LOSO CV to simulate the test scenario of predicting on unseen surveys:

Table 4: LOSO Cross-Validation Results

Model	CV Mean	CV Std	Weighted CV	Test Score
LightGBM (seed 42)	10.18	0.91	10.17	-
XGBoost (seed 42)	10.15	0.97	10.13	-
CatBoost (seed 42)	**9.14**	1.03	9.12	**9.353**
Quantile CatBoost	∼9.44	-	-	10.024
MDN (seed 0)	8.43	1.33	8.34	-
MDN (seed 1)	**7.72**	0.82	7.75	**12.038**
MDN (seed 2)	8.40	1.57	8.35	-
MDN (seed 3)	9.85	0.54	9.84	-
MDN (seed 4)	9.37	1.71	9.32	-
Competition Leader	-	-	-	**3.207**

5.2 Feature Importance

The top features across models consistently included:

Table 5: Top 10 Features by Importance

Rank	Feature	Description
1	log_utl_exp_per_person	Log utility expense per capita
2	utl_exp_per_person	Utility expense per capita
3	sfworkershh	Share of formal workers
4	educ_max	Maximum education level
5	strata	Survey stratum
6	food_diversity	Number of food items consumed
7	hsize_squared	Squared household size
8	food_diversity_ratio	Food diversity normalized
9	any_formal_worker	Has formal worker indicator
10	high_strata	High stratum indicator

5.3 Submission History

Table 6: Competition Submission Results

Submission	Method	CV Score	Test Score	Rank
Initial	CatBoost Ensemble	9.14	9.353	#211
Quantile	CatBoost Quantile	∼9.44	10.024	-
Neural	MDN (5-component)	7.72	12.038	-
Leader	Unknown	-	3.207	#1

6 Discussion

6.1 The CV-Test Generalization Gap

The most striking finding is the poor generalization from CV to test, particularly for the MDN approach which had the best CV score (7.72) but worst test score (12.04). Several factors may explain this:

1. **Distribution shift**: The test surveys may have different characteristics than training surveys. With only 3 training surveys, the CV cannot capture all possible variations.
2. **Overfitting to training distribution**: The MDN's ability to model complex distributions may have led to overfitting the specific patterns in training data that don't generalize.
3. **Calibration issues**: The MDN may have learned biased variance estimates, leading to systematically wrong poverty rate predictions.
4. **High variance of LOSO**: With only 3 folds, LOSO CV has high variance and may not be reliable for model selection.

6.2 Why Simple Models Performed Better

The CatBoost point prediction approach, despite not explicitly modeling distributions, performed more robustly:

- Strong regularization (early stopping, L2 penalty) prevented overfitting
- Log-transformed target naturally handles the skewed consumption distribution
- Simpler model has less capacity to memorize training-specific patterns

6.3 The 3x Gap to Competition Leader

The competition leader achieved 3.207, roughly 3x better than our best submission (9.35). Possible approaches that might explain this gap:

1. **Better calibration**: Direct optimization of poverty rates rather than consumption
2. **External data**: Using auxiliary datasets to improve generalization
3. **Survey-specific modeling**: Learning survey-specific biases and corrections
4. **Ensemble of diverse approaches**: Combining many different model types
5. **Better understanding of the domain**: Expert knowledge about poverty measurement

6.4 Lessons Learned

1. **CV doesn't always predict test performance**: Especially with limited validation data, CV may be misleading
2. **Simpler is often better**: More complex models (MDN) can overfit even with proper regularization
3. **The metric matters**: Optimizing consumption MAPE doesn't directly optimize poverty rate accuracy
4. **Domain knowledge is crucial**: Understanding how poverty rates are measured and what drives their variation is essential

7 Conclusions

We presented a comprehensive analysis of methods for household consumption imputation and poverty rate prediction. Our key findings are:

1. Gradient boosting (CatBoost) provides a strong baseline with CV score 9.14 and test score 9.35
2. Quantile regression offers interpretable distribution modeling but didn't improve test performance (10.02)
3. Mixture Density Networks achieved the best CV score (7.72) but worst test performance (12.04), highlighting the generalization challenge
4. A significant gap remains to the competition leader (3.21), suggesting fundamentally different approaches may be needed

7.1 Future Directions

1. **Direct rate optimization**: Train models to directly minimize poverty rate error
2. **Robust estimation**: Use methods designed for distribution shift scenarios
3. **Semi-supervised learning**: Leverage test features to improve predictions
4. **Bayesian approaches**: Properly quantify uncertainty in poverty estimates
5. **Multi-task learning**: Jointly predict consumption and poverty indicators

Acknowledgments

This research was conducted as part of a poverty prediction competition designed to explore imputation methods for real-time poverty monitoring.

References

[1] Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NeurIPS*.

[2] Prokhorenkova, L., et al. (2018). CatBoost: unbiased boosting with categorical features. *NeurIPS*.

[3] Bishop, C. M. (1994). Mixture Density Networks. Technical Report, Aston University.

[4] Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 33-50.

[5] World Bank. (2020). Poverty and Shared Prosperity: Reversals of Fortune.