Machine Learning

Final Project

Aggelos R. Dionysatos (246)

Introduction

Measuring poverty is essential for designing effective public policies and allocating resources efficiently. However, many countries lack frequent, reliable and consistent household-level surveys due to economic constraints, conflicts or political instability. As a result, machine learning models that can estimate welfare indicators from limited observable data become extremely valuable tools for governments and international organizations.

The goal of this project is to predict household consumption levels, which are used as a proxy for poverty, using survey-based household features provided by the World Band and DrivenData competition platform. This is formulated as a regression problem, where the target variable is continuous household consumption.

The objectives of this project are:

- To preprocess and analyse real-world socio-economic data.
- To explore relationships between household characteristics and poverty
- To apply and compare multiple machine learning and deep learning models
- To evaluate models using validation techniques and leaderboard feedback.
- To interpret results and discuss limitations and future improvements

Dataset Description

The dataset is provided by the World Bank through DrivenData and contains household-level survey data. The main training files include:

- train_hh_features.csv: contains the Household features
- train_hh_gt.csv: contains the target variable for household consumption
- test_hh_features.csv: test data used for prediction and submission

Feature Categories

The features cover multiple socio-economic dimensions:

- Identifiers and sampling information, which include household IDs and survey weights that represent population-adjusted sampling probabilities
- Welfare and expenditure indicators, to capture household spending behaviour
- Demographics and household composition, such as household size and age distribution
- Education and employment
- Housing and utilities, such as electricity access, water source and sanitation
- Food consumption over the last seven days, which can correlate with welfare.

Data Preprocessing

Real-world survey data contains missing values, redundant features and mixed data types, which require careful preprocessing.

Data Merging and Cleaning

- Household features were merged with consumption labels using household IDs
- Non-informative identifiers were removed.
- Columns with excessive missing values were either dropped or imputed.

Handling Missing Values

Different strategies were applied depending on feature type:

- **Numerical features:** Missing values were filled using median imputation.
- **Categorical-like features:** Missing values were filled using mode or treated as a separate category (especially for CatBoost, which can natively handle missing values).

Feature Scaling

For neural network training, numerical features were standardized using **StandardScaler**, which improves gradient stability and convergence speed. Tree-based models do not require scaling.

Train-Validation Split

Since the test set does not contain labels, internal validation was required:

- The dataset was split into training and validation sets using an 80/20 ration.
- Evaluation was performed using **Root Mean Squared Error (RMSE)**, consistent with the competition metric.

Data Analysis

Feature Distributions

Several variables showed skewed distributions, especially expenditure-related and food consumption variables. This is expected in socio-economic data, where many households cluster near lower consumption levels, with fewer high-consumption households.

Correlations

Correlation analysis showed:

- Strong positive correlations between education indicators and consumption.
- Housing quality and access to utilities were strongly associated with higher welfare.

Feature Importance

Feature importance was later extracted from tree-based models and showed consistent patterns:

- Education and employment indicators were among the strongest predictors
- Food consumption variables were highly informative
- Housing conditions significantly contributed to prediction quality

This confirms that poverty is multi-dimensional and cannot be inferred from a single category of features.

Machine Learning Models

LightGBM Regressor

LightGBM is a gradient boosting framework optimized for large datasets and high-dimensional features. It was selected for its strong performance in tabular regression tasks and is widely used in Kaggle/DrivenData competitions.

Advantages:

- Fast Training
- Handles missing values
- Captures complex non-linear relationships

XGBoost

XGBoost is another gradient boosting method known for its robustness and regularization mechanisms. It was selected as a reliable baseline model and strong competitor to LightGBM

Advantages:

- Prevents overfitting with built-in regularization
- Good performance with heterogeneous features

CatBoost Regressor

CatBoost is specifically designed to handle categorical variables effectively. It is well-suited for socio-economic survey data with mixed variable types.

Advantages:

- Native handling of categorical and missing data
- Reduces preprocessing requirements

Neural Network (Deep Learning)

A fully connected feedforward neural network was implemented using PyTorch. This was used to evaluate whether deep learning provides benefits over tree-based methods for tabular data.

Architecture:

- Input layer equal to number of features
- Two hidden layers with ReLU activations
- Output layer predicting continuous consumption

Training:

- Optimizer: Adam
- Loss function: Mean Squared Error (MSE)
- Data normalization before training

Model Evaluation and Comparison

Validation Strategy

Since test labels are not available, performance was evaluated using Holdout validation set and RMSE metric.

Results Summary

The ensemble and individual models achieved approximately:

- **Best Individual Model RMSE:** 5.89
- **Ensemble RMSE:** 5.86
- **Relative Improvement:** ~0.6%

Although the improvement is small, the ensemble consistently outperformed each individual model.

Ensemble Method

Predictions from LightGBM, XGBoost and CatBoost were averaged.

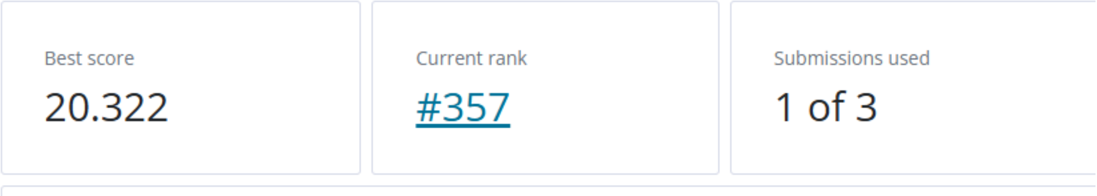
Ensemble works because:

- Each model learns different decision boundaries
- Errors made by one model may be compensated by others.
- Averaging reduces variance and improves generalization

This is especially effective in complex, noisy socio-economic datasets

Leaderboard Evaluation

Predictions were generated for the test dataset and submitted to the DrivenData platform using the required submission format.



Best score	Current rank	Submissions used
20.322	#357	1 of 3

Interpretation of Results

Feature Influence

Tree-based feature importance analysis showed that models focused primarily on:

- Education
- Employment
- Food consumption metrics
- Housing and utilities access

These factors confirm that poverty is influenced by structural living conditions rather than isolated variables.

Where Models Perform Well

Models perform well when:

- Household survey responses are complete
- Living conditions are clearly indicative of poverty or affluence

Where Models Perform Poorly

Limitations appear when:

- Household exhibits unusual consumption behaviour
- Regional economic conditions differ from training distributions
- Informal income sources are not captured by surveys.

Possible Improvements

Future improvements could include:

- Applying country-specific fine-tuning models
- Using more advance ensembles such as stacking.

Conclusion

This project demonstrated how machine learning and deep learning can be applied to real-world poverty estimation problems using household survey data. After preprocessing and feature analysis, multiple regression models were trained and compared. Tree-based ensemble methods achieved the best performance and combining them further improved prediction accuracy.

The results highlight the importance of combining diverse socio-economic indicators when estimating welfare and confirm that ensemble learning is an effective strategy for noisy, high-dimensional survey datasets.