TOLGA ENGIN 204

Poverty Prediction Using Machine Learning and Deep Learning

World Bank Poverty Measurement Challenge

1. Introduction & Problem Description

Poverty measurements are critical for countries and its people as well as governments. It has strong influence on how the policies are made and decided. In traditional poverty assessment, the method was to go to each single individual household and survey them. However this whole process is expensive, takes a lot of time and resource heavy.

The goal of this project is to be able to predict the per capita household consumption. This will be done using survey based household features. Also another goal of this whole project is to derive poverty distributions across predefined consumption thresholds. This project is based on a real world dataset. The dataset is provided by the World Bank. The results will be evaluated by the platform DrivenData.

The project also targets to establish an end to end machine learning pipeline. This whole process will also include data preprocessing, exploratory analysis, model training, deep learning, and external validation. The submission will be done to DrivenData platform and the results will be on a public leaderboard.

2. Dataset Description

The dataset consists of three main files:

- **Training household features**: This dataset represents the socio-economic and demographic variables at household level
- **Training labels**: This dataset is dedicated for per-capita household consumption (cons_ppp17, PPP-adjusted USD, 2017)
- **Validation (test) household features**: This features without labels, used for final evaluation

In all these, each household is uniquely identified by the following attributes:

- survey_id
- hhid

The target variable is continuous. As it is continuous it makes this a regression problem.

3. Data Preprocessing & Feature Engineering

Data preprocessing was implemented. This will ensure that the data is clean and understandable to use. It will ensure consistency among the whole pipeline. In this step data merging took place. Missing valued were handled with elegance. Categorical variables were encoded using One-Hot Encoding,

4. Exploratory Data Analysis (EDA)

Exploratory analysis focused on understanding the target variable and baseline model behavior.

5. Machine Learning Models

To be able to establish the machine learning pipeline different model were used in this whole assignment. These models include

-Ridge Regression

-Random Forest Regressor

-Gradient Boositng/XGBoost

6. Deep Learning Model

A deep learning approach was also implemented to meet the project requirements. Dimensinality reduction was applied. Duıring the training of the Neural Network Architectıre the following techniques were implemented:

-Two hidden layers

-ReLU activations

-Dropout for regularization

-Mean Squared Error loss

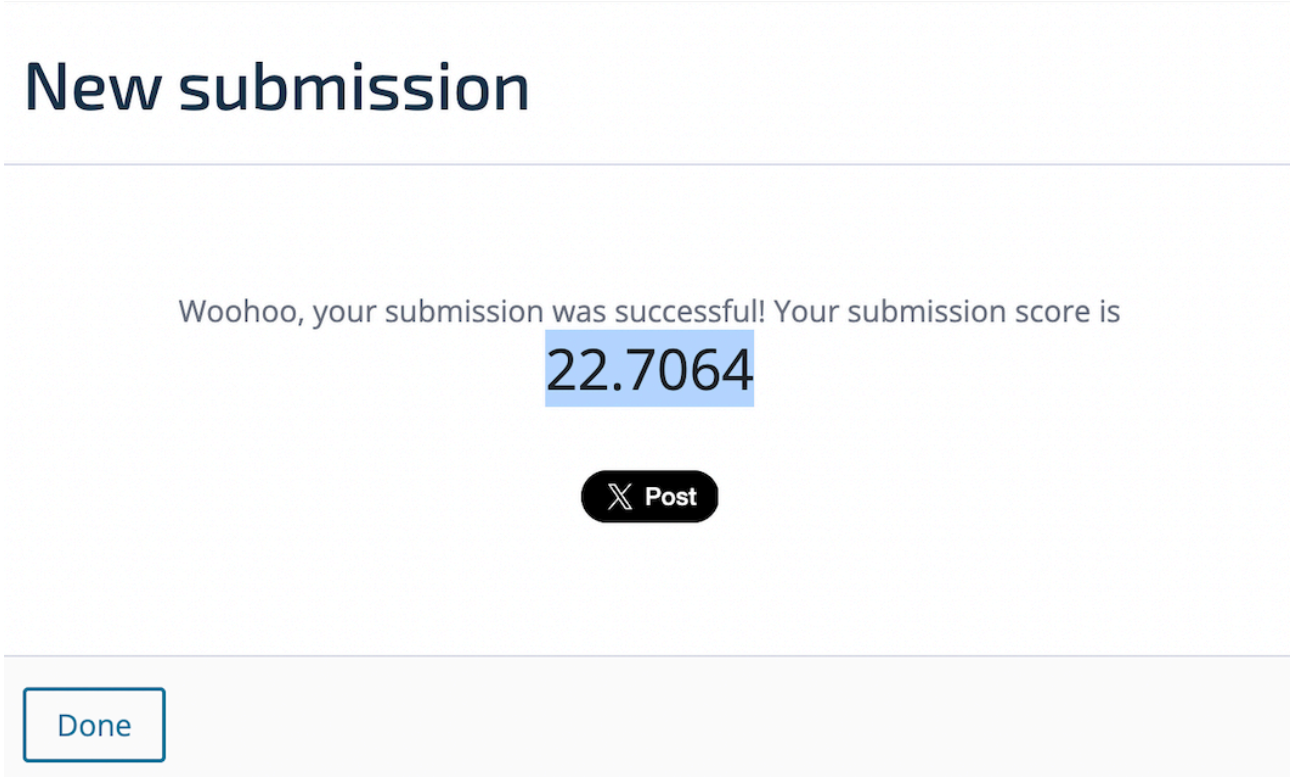
7. Poverty Distribution Estimation

Other than household level consumption prediction, the competition requires estimating poverty distributions. Using predicted household consumption, poverty rates were

computed for 19 predefined thresholds. With this step, individual household predictions were transformed into policy relevant poverty indicators.

8. Validation & Submission

The final model was evaluated externally by submitting predictions to the DrivenData platform. The result of the submission can be seen in the image below.



9. Conclusion

In this project, an end-to-end machine learning pipeline was developed and deployed. This pipeline was made for poverty prediction using real world World Bank dataset. The workflow included data preprocessing, exploratory analysis, classical machine learning, deep learning, and external validation.

The results demonstrate that model performed particularly well on structured socio economic data, while deep learning provides a complementary approach. The project highlights the challenges and practical considerations in real world applied machine learning.