



Machine Learning Approach

Targeting the Poverty in Peru

Rachid Ben Maatoug
Nataell Cornu
Mingmin Feng
Jingyan Yang
Yu Pan

Department of Economics

2020/12/16



Raw Data Overview

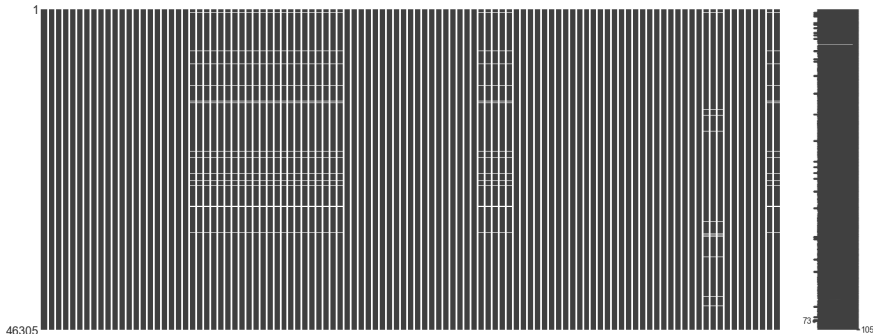
- Training dataset (23,153 Obs.) & test (23,152 Obs.)
- Binary inputs & continuous output
- Use regression algorithm

Build Derived Features

- 5 asset variables:
internet, cable, computer, refrigerator, washer
- 2 time variables: *year* and *month*
- 1 geographical variable: *dominio*

Handle Missing Values

- Training: delete the records
- Test: impute with Median





ML Approach Overview

- Cross-validation
- Metric: MSE
- Approaches:
 - Ordinary Least Square (OLS) : baseline
 - K-Nearest Neighbors(KNN)
 - Lasso
 - Support Vector Machine (SVM Regression)
 - Light Gradient Boosting Machine (Light GBM)
 - Neural Network: Multi-Layer Perception (MLP)

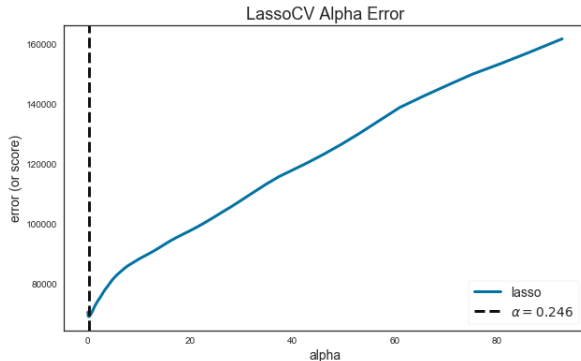
Summary Statistics

Table 1: Overview of MSE

Model	MSE.y	MSE.lny
OLS:	89689.4794	0.1909
Lasso:	80288.0317	0.1805
KNN:	90780.4793	0.2099
SVR:	84657.0160	0.1792
Light GBM:	80546.3523	0.1786
NN:	81377.0992	0.1870

Lasso

- Generate 4560 interaction terms
- $MSE = 0.1805$





SVM and Light GBM

- SVM
 - Model nonlinear relationship
 - Computational expensive
 - $MSE = 0.1792$
- Light GBM
 - Ensemble method
 - $MSE = 0.1786$

Figure 5 Replication(Overview $\rho = 3$)

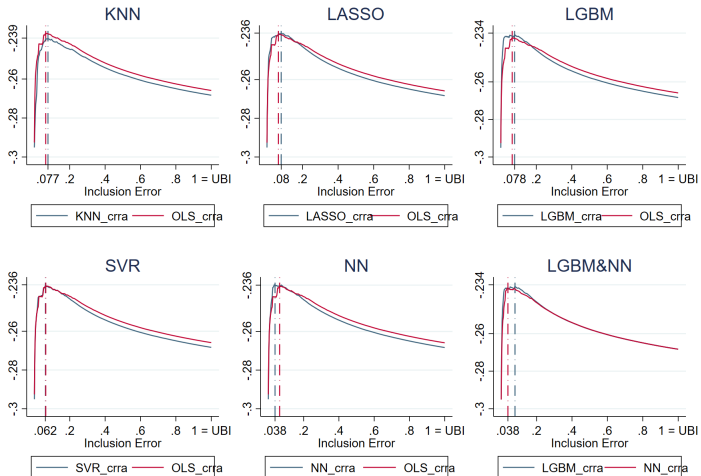
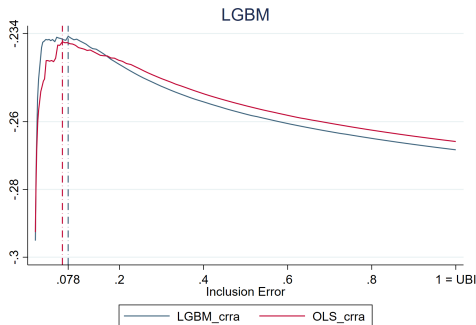


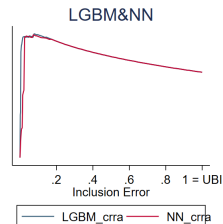
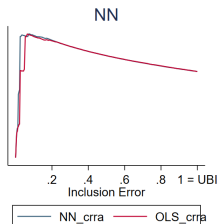
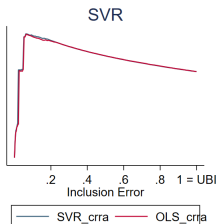
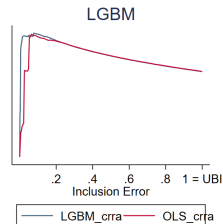
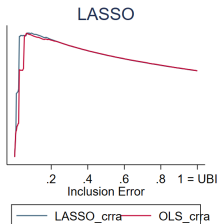
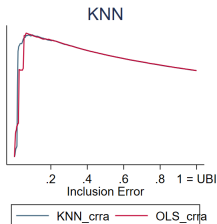
Figure 5 Replication(OLS vs Light GBM)



	OLS	Light GBM
CRRA utility	-0.2364	-0.2348
Targeted population	18%	21%
Inclusion error	6.4%	7.8%
Exclusion error	52.40%	45.00%

Robustness Check: $\rho=5$

- Change the weights on the poor and the rich





Thank you!