Abstract

The world bank is an international financial institution that develops poverty reduction strategies and goals. DrivenData released a data analysis competition in collaboration with The World Bank to predict the probability of poverty in households in developing countries. The intent of the competition was for the World Bank Development Data Group to engage data scientists to develop effective poverty prediction methods.

The data released consisted of 6 datasets: two for countries A,B and C. Two datasets for each country consisted of one for individuals, and one for households. Each variable in the datasets were random strings. The datasets differed between the three countries, meaning that each analysis had to be tailored to each dataset specifically. Every dataset contains a ‘poor’ variable with two levels. Dimensions of each dataset are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Grouping | Variables | Observations |
| A | Household | 346 | 8203 |
| A | Individual | 44 | 37560 |
| B | Household | 443 | 3255 |
| B | Individual | 227 | 20252 |
| C | Household | 165 | 6469 |
| C | Individual | 44 | 29913 |

The analysis consisted data preprocessing (involving joining the individual and household datasets, removing outliers, dealing with NAs) before data could be modelled. Additionally, the high dimensional nature of the data required feature engineering. The primary ways of doing this were to remove variables which caused problems with the analysis (variables with too many levels, or NAs) and additionally variables which had little variability. Additionally, recursive feature elimination was used.

Predictive algorithms used included KNN, Logistic Regression Random Forest, GBM and XGBoost. XGBoost currently yields the highest accuracy. Complications with modelling the data included overfitting.