# Machine learning techniques for measuring poverty

Suchit Gabrani

Poverty elimination is one of the noblest goals of our generation. A first step of poverty reduction and elimination is finding which parts of the world the most poor live, so that the necessary help can be provided. Understanding of poverty prevalence is important to ensure that Government and NonProfit resources can be allocated appropriately.

Machine learning has been used in the last few years to solve this challenge, with great success. In this review we will look at a few recent articles and papers in the field of using Machine Learning for estimating poverty. We will analyze the machine learning algorithms that were used. The focus of the review will be on the critique on machine learning algorithms used for identifying poverty laden areas.

Data gaps in developing countries including African and South Asian countries have impacted the planning and implementation of poverty reduction programs. Closing the data gaps through surveys can be quite expensive. Novel data sources used along with creative Machine Learning techniques can play a big role here.

We look at different approaches used to predict poverty using machine learning.

**Applying modern machine learning technique for improving poverty estimating surveys**
Doing household surveys is the traditional method of estimating poverty of a region. This can also be improved by using modern machine learning techniques like cross-validation and parameter regularization.

In order to accurately identify poor households, assessment via surveys can take several hours per household. This can be expensive and cumbersome. Proxy-means test (PMT) [8] uses a small number of questions (say 10 to 30) to estimate the probability of a household being poor. The survey results are fed into a scorecard to get a final result without much calculations [3]. As a simple example a question may be  “Does every child between 6 and 12 attend school?”, if the answer is yes, 100 points are added to the score. If the answer is no then 0 points are added to the score, and 50 points are added to the score for other answers.

Kshirsagar et al [4] used a novel PMT generation approach using modern machine learning techniques like variable selection and regression. The technique used satisfies three important criteria:
i) The prediction model uses ten questions. This leads to lower costs of data collection;
ii) Computation to calculate the probability that classifies a household as poor or not, uses simple arithmetic and can be done immediately after data collection
iii) One scorecard (model) is used to predict poverty throughout a country.

The technique consists of the following steps:

1. Variable selection: The survey chooses the set of ten questions that have been answered most frequently in the training data. This allows a model that works well with diverse parts of a country.
2. Fitting the selected model: For the ten variables selected in the first step, a new survey-weighted elastic net logistic regression is computed for all observations in the training data. Here, α is selected using an outer cross-validation loop over this. λ is chosen using internal cross-validation.
3. Translating the model into scorecard format: we shift, scale, and round The estimated logistic regression coefficients $\hat{\beta}$ is shifted, scaled and rounded into weights $\tilde{\beta}$, such that all combinations that are valid. of the linear predictors that have an integer scores in between 0 and 100.

As an example the survey for Zambia includes a question: “What material was used to construct the roof?” For this question if the solution Thatched indicates a poor household, and corresponds to zero points. The household gets 7 points for Iron Sheets, and 10 points for Cement/ Asbestos/Other, as this indicates the household is not poor. The survey agent then adds the score of ten questions. If that is below a threshold value, then the household is classified as poor.

The advantage of this technique is that a computational intensive process is used to calculate the model. However the actual classification of households to poor/not-poor involve very simple calculations.

**Predicting using daylight satellite images using transfer learning models**

Satellite high resolution imagery about landscape features is available for different parts of the world. However not much labeled data is available. Machine learning techniques like deep learning are effective when abundant data is available [2]. Due to the scarcity of training data, deep learning models such as convolutional neural networks can not be applied to the high resolution satellite images about landscape features.

Jean, N et all have studied the use of daylight images to predict poverty levels with impressive results [1]. They perform analysis on the data of five Aftican countries: Uganda, Nigeria, Tanzania, Rwanda and Malawi. The same approach can be applied to other geographies.

Convolution Neural Network (CNN) has become the most popular tool used in the computer vision community to achieve feature extraction. Estimating economic wellbeing from satellite imagery is non-trivial. So labelled training data for satellite imagery is scarce. To address this the authors solve a proxy problem[1]. Imagenet provides a dataset of millions of labeled training examples. To address this, they use an eight layer CNN model previously trained on ImageNet dataset, and use that as training data to estimate nighttime light intensity at various locations based on the corresponding daytime satellite images. The nightlights data is used to get an intermediate label to learn image features. These image features are related to economic well-being. This step is considered as a classification problem. In the final step the model is

directly used to estimate local per capita outcomes from daytime image features. This does not rely on nightlights.

Satellite night data is available at 1 KM resolution. This is mapped to one of the three light intensities: low, medium and high. The model is trained using minibatch gradient descent. For each household cluster there are 100 input images for 10Km by 10Km area. For each cluster 100 feature vectors are generated which are then averaged to get one feature vector.

The CNN model parameters are represented as filters and can look for features like roads and buildings then these are used to predict poverty. They demonstrate that using the existing high-resolution day time satellite images may be used to make fairly accurate predictions about the prevalent poverty in African countries. Similar approaches can be applied to other developing countries as well. The limitation of the model is that it is not able to distinguish between the households within a cluster, and all households in a cluster are put in the same category.

Table 1 displays the $r^2$ of the Predicted consumption with Observed consumptions[1]. It indicates that the satellite imagery gives good results in predicting poverty levels

| Country | $r^2$ |
|---|---|
| Nigeria | 0.42 |
| Tanzania | 0.55 |
| Uganda | 0.41 |
| Malawi | 0.37 |

**Table 1**: Predicted cluster-level consumption from transfer learning approach compared to survey-measured consumption

**Comparison of performance of Machine Learning Approaches for evaluating poverty surveys**

Sani et al[5] have analyzed the problem of Bottom 40 Percent Households [B40] using machine learning techniques. In the next decade Malaysia is expected to enter the high-income economy status if the B40 households could be helped by Government programs

The authors compare the performance of Naive Bayes, Decision Tree and k-Nearest Neighbors algorithms in computing the B40 households of Machine Learning. They look at the dataset called eKasih consisting of 99,546 households from 3 different states to train different machine learning models. The results of the models are compared through a 10-Fold Cross-Validation method.

The authors follow a three step approach:

1. Dataset phase: The dataset phase consists of identifying data to be examined in this study and analysing its details. ‘eKasih’ database keeps detailed profiling of the poor

households in Malaysia including parameters like: Ethnicity, Marital status, age, gender, type of job, education degree, house ownership type. The citizens are classified into: Poor, Hardcore Poor and Excluded.

2. Preprocessing of data to get it ready for application of algorithms. The data from the real world is often incomplete, and it may contain errors and outliers. It needs to be preprocessed before algorithms can be applied. The following tasks are performed on the data: data cleaning, feature engineering, Normalisation, and sampling techniques using a tool called 'Waikato Environment for Knowledge Analysis (Weka)' version 3.8 Software [6,7]. Weka also provides data visualization functionality. Outliers' detection is conducted using Interquartile Range in Weka. The outliers are kept, as data is deemed to be authentic. Normalization is performed on parameters like age, income to bring the data in the range 0 to 1. Three ranking methods are used for feature extraction: Correlation Attribute, Information Gain Attribute and Symmetrical Uncertainty Attributes

3. Application of three machine learning techniques on the dataset: Three popular machine learning classification techniques are applied on the preprocessed data Naïve Bayes, Decision Tree and k-Nearest Neighbors. Each of these techniques are tuned using appropriate parameters. The Naïve Bayes is tuned using Discretization. Discretization is defined as transformation of numeric data into nominal data. The numeric values are organized into distinct groups of fixed length. In the Decision tree model, an important tuning parameter is the confidence factor. This parameter determines whether an attribute with a certain value belongs to a certain class or not. Another decision tree parameter is the threshold of the minimum number of instances at the leaf node, beyond which the parent node and its children node are compressed into one node. The k-Nearest Neighbors use the following tuning parameters: k-Value (number of nearest neighbors), and the distance functions (Euclidean distance, Chebyshev distance, Manhattan distance and Minkowski distance).

   Two techniques were used by the authors [5] to determine the Classification Accuracy and Kappa Statistic. Classification Accuracy, expressed in percentage is the ratio of the number of correct predictions to total predictions. Kappa Statistic, expressed between 0 to 1, captures the similarity between two classifications. The experiment is validated using Cross-Validation (10-Fold ). The results using the top 8 attributes show more than 90% Classification Accuracy using all the three algorithms. The best results were achieved using Decision Tree (99.27%), followed by Naïve Bayes (97.27%) and k-Nearest Neighbor (96.8%).

Table 2 demonstrates the results achieved on 10-fold cross-validation by different algorithms achieved by Sani et al[5].

| Algorithm used | 10-fold cross-validation results |
|---|---|
| Naïve Bayes | 97.27% |

| | |
|---|---|
| k-Nearest Neighbor | 96.8% |
| Decision Tree | 99.27% |

**Table 2**: 10-fold cross-validation results

# Conclusion

Tables 3 through 5 includes a summary of techniques discussed in this survey.

| | |
|---|---|
| Researchers | Kshirsagar, V., Wieczorek, J., Ramanathan, S. and Wells, R[4] |
| Geographies | Zambia |
| Data Analyzed | Proxy-means test: Very simple Household surveys with ten questions |
| Algorithms used | Elastic net logistic regression, Cross-validation |
| Summary | The model development which involves giving weights to different answers to the survey questions involves fairly complex mathematical models. Surveys simple, and household classification calculations to poor/not-poor involves very simple calculations. So time to conduct surveys is significantly reduced |

**Table 3**: Improving Proxy-means test using Machine Learning

| | |
|---|---|
| Researchers | Jean, N., Burke, M., Xie, M., Davis, W., Lobell, D. and Ermon, S [1] |
| Geographies | Uganda, Nigeria, Tanzania, Rwanda and Malawi |
| Data Analyzed | Night-time satellite imagery along with Imagenet dataset |
| Algorithms used | Convolution Neural Network, CNN, Minibatch Gradient descent |
| Summary | Satellite night data is available at 1 KM resolution. This is mapped to three light intensities: low, medium and high. The model is trained using minibatch gradient descent. The Convolution Neural Network, CNN model parameters are represented as filters. They are used to identify features like roads and buildings. These are then used to predict poverty. |

**Table 4**: Using Night-time satellite images along with Imagenet data to predict poverty

| | |
|---|---|
| Researchers | Sani N. S., Rahman, M.A., Bakar, A.A., Sahran, S. and Sarim, H.M [5] |
| Geographies | Malaysia |

| | |
|---|---|
| Data Analyzed | Comprehensive household data |
| Algorithms used | Data cleaning, feature extraction, Normalisation, and sampling techniques. Naïve Bayes, Decision Tree and k-Nearest Neighbors |
| Summary | Detailed household data including Ethnicity, Marital status, age, gender, type of job, education degree, house ownership type for Malaysia is available. The citizens are classified into: Poor, Hardcore Poor and Excluded. Three popular machine learning techniques are applied/compared and very high accuracy in prediction is achieved. |

**Table 5**: Bottom 40 Percent Households using Machine Learning

Machine learning techniques in measuring poverty have been used to expedite and reduce expenditure of traditional surveys techniques for poverty measurement. Also original techniques like measuring lighting via satellites are being used as a low cost measure for detecting poverty. These techniques allow using donations for alleviating poverty rather than just measuring it. These techniques are still in their infancy. Several other sources may be used to predict poverty like cell phone usage, purchase patterns, internet usage. The next decade will see machine learning techniques usage to predict poverty becoming more widespread.

# References and Notes


1. Jean, N., Burke, M., Xie, M., Davis, W., Lobell, D. and Ermon, S. 2016. “Combining Satellite Imagery And Machine Learning To Predict Poverty.”, Science vol. 353(6301), pp. 790-794
2. LeCun,Y., Bengio,Y. and Hinton,G. 2015. “Deep learning.”, Nature vol. 521, pp.436–444 . doi:10.1038/nature14539pmid:26017442
3. Schreiner, M. 2007 “A simple poverty scorecard for the Philippines.” Philippine J. of Development vol 34(2), pp.43-70. https://dirp3.pids.gov.ph/ris/pjd/pidspjd07-2poverty.pdf
4. Kshirsagar, V., Wieczorek, J., Ramanathan, S. and Wells, R., 2020. “Household Poverty Classification In Data-Scarce Environments: A Machine Learning Approach.” arXiv.org. <https://arxiv.org/abs/1711.06813>
5. Sani N. S., Rahman, M.A., Bakar, A.A., Sahran, S. and Sarim, H.M. 2018 “Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification,” International Journal on Advanced Science, Engineering and Information Technology, vol. 8(4-2), pp. 1698.
6. Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. 2016. “Data Mining: Practical machine learning tools and techniques.”4th Ed. Morgan Kaufmann.
7. WEKA,” Weka 3 - Data Mining with Open Source Machine Learning Software in Java. <https://www.cs.waikato.ac.nz/ml/weka/>

8. Worldbank Group.”Measuring income and poverty using Proxy Means Tests.” <https://olc.worldbank.org/sites/default/files/1.pdf>