POVERTY ANALYSIS, PREDICTION USING MACHINE LEARNING METHODS

Prof. V. D. Ghonge, Parth Sandeep Kadam, Vitthal Arjun Bhakare, Amit Kailas Bodake Prashant Sudhakar Warungase

Department of information technology, Smt. Kashibai Navale College of Engineering. Maharashtra, Pune -411041 India

ABSTRACT:

Lack of sufficient resources to provide for basic needs like food, clean water, housing, and clothing, as well as in today's world, access to health care, education, and even transportation, is referred to as poverty. The government of the country was given many methods, but they don't function as they should. The predictions are inaccurate, and the country's customary method of making predictions involves conducting a site survey, which is pricy and labor-intensive and sometimes a waste of time before learning the actual outcome. Making educated policy decisions and efficiently distributing resources to the places that need assistance the most is severely hampered by the absence of credible statistics on poverty in the nation. In order to fully understand the causes of poverty, we will first conduct a multidimensional analysis of poverty using multiple correspondence analysis. Next, we will make predictions using three different machine learning techniques, and finally, based on prior research, we will also use satellite images processed through convolutional neural networks to estimate the level of poverty. In order to determine whether the method is better suited to comprehend and predict poverty in a country, this paper compares simple machine learning methods to advanced deep learning methods in an effort to build on prior research.

Keywords: Poverty Prediction, Machine Learning, Algorithms, datasets, Prediction

==========

[1] INTRODUCTION

Poverty is a major issue in our nation with multiple dimensions and multiple classifications. For some authors, it is determined by money, while other researchers additionally take into account factors including health, education, social standing, and political rights. But, what unites these scholars is their work in the areas of identifying causes that generate poverty, classifying people based on various perspectives of poverty, and forecasting future levels of poverty. [1]. our goal

in writing this paper is to use a household survey dataset to address the issue of poverty in the country. We then use three straightforward poverty prediction models and satellite images of two different states—Lagos and Jigawa—to develop a sophisticated convolutional neural network model. We will then compare these models to determine which best captures poverty in the country. We will use logistic regression, decision trees, and random forests as machine-learning techniques to build our models. There are sections in our paper. We shall discuss multidimensional poverty analysis in the first section. The second section discusses the techniques we employed and describes our process. After presenting our observations and outcomes in the third section, we draw conclusions in the fourth section.

2. LTERATURE REVIEW

We will briefly discuss the study's literature review in this chapter. We'll start out by talking about multidimensional poverty analysis and poverty measurement. After that, we'll describe the theoretical underpinnings of the statistical approaches used to perform the poverty analysis. Finally, we'll examine the theoretical underpinnings of the machine learning and deep learning techniques we employed in this work. Further information about the various techniques and algorithms we employ is provided in this section.

A study was conducted to compare the performance of Naïve Bayes, Decision Tree, and k-Nearest Neighbors in classifying the B40 population in Malaysia [1]. The study made use of the 'eKasih' dataset from the National Poverty Data Bank. It includes a thorough profile of Malaysia's low-income households. The importance of this work highlights feature engineering, normalizing, sampling technique selection, feature selection approaches, and parameter tweaking. To balance out the dataset, a technique known as Synthetic Minority Oversampling Technique (SMOTE) is used to construct replica cases. The different combinations of parameters used to optimize each classifier include discretization for Naive Bayes, confidence factor, and the minimum number of objects for Decision Trees, as well as k-value and distance function for k-Nearest Neighbors. By ranking the top eight features utilizing symmetrical uncertainty, correlation, and information gain attributes, feature selection algorithms have been found to increase classification accuracy and the Kappa statistic. K-Fold Cross-Validation of 10 is used as the measurement for assessing the performance of all three classifiers after parameter tuning, and a statistical test is run to determine whether two "models are statistically significantly different from one another or if one of them is better than the other." [1]. the study concludes that the Decision Tree model is strictly outstanding and has surpassed other classifiers in respect of accuracy.

Furthermore, in [2] a general population of Hong Kong, the neighborhood-level and individual-level factors of poverty were investigated. Prior research mostly concentrated on using financial indicators to evaluate poverty and poverty within a certain population. Yet, this study strongly emphasizes a holistic approach to address all aspects of what determines poverty. [2], which makes use of the lack that impoverished people suffer, such as insufficient income, bad health, and lack of knowledge. The author used Quantile Regression to further analyze the differences in the effects of the determinants across five poverty spectrums after using Logistic Regression to study the determinants of poverty. When the poverty line or threshold is employed as the measurement, logistic regression is typically used to assess the level of poverty. Only the percentage of persons who live in poverty may be determined using this method. It does not provide a variety of explanations for the experiences of the impoverished. A more thorough explanation of poverty status is provided by Quantile Regression, which identifies the differential outcomes of the factors that determine poverty across the poverty spectrum. Based on the ratio of income to poverty (I/P), quantile regression calculates the level of poverty. This study defines six poverty thresholds based on the poverty line for Hong Kong in 2015. These thresholds apply to households with up to six people. Also, a specific amount of the I/P ratio is mapped to five quantiles. Each quantile is mapped to a particular category of poverty status. The

quantile regression model's findings provide magnitudes of connections between different factors and poverty status, including whether a given variable is significant or not across the range of poverty and whether it is positively or negatively linked with poverty. In order to compare "how some quantiles of the I/P ratio may be more affected by a given predictor than other quantiles," [2] Ordinary Least Square (OLS) regression is used in statistical analysis. The use of cluster-robust standard errors and data weighting for the oversampled data is also described. Based on household income and using satellite photos of the urban environment in North and South America, the task of predicting poverty was examined in [3]. The cost and labor requirements of the method used to gather socioeconomic data served as the study's driving forces [3]. As a result, it turns to remote sensing data, which is more suited for estimating poverty on a big scale, including high-resolution satellite photography. To create a descriptive urban landscape for identifying the urban areas in each city, original satellite data is combined with crowdsourced OpenStreetMap (OSM) data. Regression and convolutional feature extraction are used in the study to estimate the location of objects. For features extraction, a transfer learning procedure from three ConvNets is employed, specifically ResNet50 with initialized ImageNet weights, VGGF pre-trained on ImageNet weights, and VGGF fine-tuned with nightlight intensity from a few African nations [4]. Two separate tiers of census area boundaries are used to extract two different sorts of socioeconomic information from an input image. Before mapping, each input image is additionally rotated and flipped either horizontally or vertically. ResNet50 feature is picked as the model after all three neural networks have undergone cross-validation. In the meantime, the Ridge Regression model is used to complete the household income prediction job based on image-level characteristics and cluster-level features. The model's performance evaluation is done by using 10-fold cross-validation and regression score, metrics score.

The causes of poverty have been investigated using ordinal and multinomial logistic regression models [5]. According to this study, there are three levels of poverty: absolute poverty, near poverty, and above near poverty. Based on the household income percentage below the poverty criterion, which is set at 100% to 125% for near-poverty states and greater than 125% for states that are above near-poverty states, each state is evaluated. The threshold is multiplied by the inflation rate for each succeeding year and is based on the national median income for Poland in 2000. Also, the information is based on the number of households from 2000 to 2015, multiplied by two. In line with [6], In the Social Diagnosis 2015 study, two questionnaires were used to gather information from families. The first survey involves face-to-face interviews with the household substitute, who is the expert on the members of the home and their current situation, to gather information about the make-up of the household and living conditions. It provides a wealth of information on household composition, living conditions, and the demographic and socioeconomic circumstances of each household member. All household members who are 16 years of age or older are asked to complete the second survey. It focuses on issues that reveal a person's level of well-being. Gender, age, education level, place of residence, number of household members, biological family type (e.g., single with no children, couple with children, etc.), socioeconomic group, labor-force status, and presence of a disabled person in the household are the variables used to determine the factor of state of poverty during analysis [5]. According to the findings, the multinomial logit model performs better in predicting the level of poverty. Because the ordinal logistic regression model does not satisfy the requirement of parallel lines, the results may be misinterpreted. The study also notes that the variables (education, residence, employment position, and socioeconomic group) are the most important influences on the level of poverty [5].

[3] METHODOLOGY

In this section we will describe the data used for the work, how we obtained and pre-processed it, and then we will present an overview of the approach used to achieve the goal of this work.

3.1 Data

Country Living Standards Survey (2018)

The World Bank's Living Standard Measurement Study (LSMS), which examined various aspects of the Country's Standard of Living in 2018–2019, provided the data used to train and score the machine learning models. All surveys were carried out by the World Bank in collaboration with the National Bureau of Statistics (NBS). The survey includes data on households and communities. Household surveys were carried out on a variety of households, and questions about the household and its members were asked. Other areas included in the study include health, education, assets, housing, employment, and income. There were several unanswered questions in the 116,321 home survey. We had 8,229 homes' worth of information after cleansing. The description of the data is presented in the table 1 below.

Variable name	Values	Frequency	Description
School	Yes	6991	Has the person attended any school
	NO	1238	
Randweng	Yes	4570	Can the person read and write in English
	NO	3659	
Randwothr	Yes	3214	Can the person read and write in any language
	NO	5015	
Healthy	Yes	806	Did the person go to see a doctor recently
	NO	7423	
Employed	Yes	564	Is the person employed
	NO	7665	
Assets	Yes	2381	Does the household have any form of assets
	NO	5848	
Savings	Yes	564	Does the household have any form of savings
	NO	7665	
Tapwater	Yes	1916	Does the household have tap water from any source
	NO	6313	
Owntoilet	Yes	3089	Does the household own a flushable toilet
	NO	5140	
Electricity	Yes	5318	Have Electricity
	NO	2911	

Table 1. Data Description

3.1.1 Machine Learning Models

After utilizing the K-means method to categorize our dataset, the next step was to build a target column and link each household to its target, or whether it is poor or not. The dataset was then divided into training and testing, with 70% of the dataset being utilized for training and 30% for testing. Three models were developed using the approaches of logistic regression, decision trees, and random forests; each model is explained below..

Logistic Regression

We fit the logistic regression method to our model with the following R syntax:

multinom(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "multinom.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)

The output or result of the models shows that it took 15.11 seconds to make the predictions, and it has a 99% specificity, 99% sensitivity, and overall accuracy of 98%, which shows the model performed very well in making predictions. After training the model, we used the test data to make predictions, which was used to test and score the performance of the model.

Decision Tree

Next, we fit our second model using the decision tree method, fitting the decision tree method in R is supported by many packages but we chose to use the rpart package and the syntax used is:

rpart(formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)

Our model didn't perform as well as the first one, so we tried to improve it by adjusting some hyperparameters using the control function, and we eventually got a better model with an accuracy of 95%. After fitting, we also made predictions using the training data, and the output of the model showed it takes 1.09 seconds to make the predictions, has 69% sensitivity and 94% specificity, with an overall accuracy of 86%.

Random Forest

We used the R package RandomForest to grow the trees for our model, the syntax used is:

randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500, mtry=if (!is.null(y) && !is.factor(y)) max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))), replace=TRUE, classwt=NULL, nodesize = if (!is.null(y) && !is.factor(y)) 5 else 1, importance=FALSE, ...)

Our random forest-based model's output reveals that it has a sensitivity of 99%, a specificity of 99%, an overall accuracy of 98%, and a prediction time of 33.34 seconds. Although being the slowest of the three, it produced the best model out of the three.

4. CONCLUSION

Although we had problems with the data we obtained, we made it fit for research and employed various functions to make it suit all the methods and algorithms used for the research. A number of actions were taken to accomplish the goal of the research, including data sourcing and preparation. As a result, we worked with categorical variables using multiple correspondence analysis to construct a poverty index, which is a numerical representation of the data gleaned from

our categorical descriptor variables. We selected categorical variables that recorded data on household circumstances, education levels, and household health. The theoretical framework that embodies our understanding of poverty as a multidimensional term served as justification for the selection of these descriptions. In order to classify our respondents into four groups, we used the K-Means technique on the recently developed numerical measure of poverty. With the addition of this new variable, predictive modeling may continue. The supervised learning algorithms category includes the models that we used. As implied by the name, these algorithms demand that the model contain both the predictors and the outcome. The process of tuning an algorithm's parameters is represented by the learning that follows.

To sum up, we assert that, although requiring a lot of processing resources, neural networks are the algorithm with the best predictive capability. When we fed the algorithms satellite imagery, this was further supported. We further asserted that combining the two approaches could result in a powerful instrument for tackling poverty. We cannot, however, assert that we have developed a suitable model for predicting poverty based on the results we obtained. To add more or better predictors, there must be some adjustments. We advise conducting additional research in this area.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank **Prof. V. D. Ghonge**, our Guide and Assistant Professor, Faculty of Information Technology for his valuable guidance and moral support in the process of preparing this paper.

REFERENCES

[1] SANI, N. S., RAHMAN, M. A., BAKAR, A. A., SAHRAN, S., & SARIM, H. M. (2018) Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2), pp.1698.

[2] PENG, C., FANG, L., WANG, J. S., LAW, Y. W., et al. (2018) Determinants of Poverty and Their Variation Across the Poverty Spectrum: Evidence from Hong Kong, a High-Income Society with a High Poverty Level. *Social Indicators Research*, 144(1), pp. 219-250.

[3] PIAGGESI, S., GAUVIN, L., TIZZONI, M., CATTUTO, C., et al. (2019) Predicting City Poverty Using Satellite Imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*, pp. 90-96

[4] JEAN, N., BURKE, M., XIE, M., DAVIS, W. M., et al. (2016) Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), pp. 790-794.

[5] SĄCZEWSKA-PIOTROWSKA, A. (2018) Determinants of the state of poverty using logistic regression. *Śląski Przegląd Statystyczny*, 16(22), pp. 55-68.

[6] PANEK, T., CZAPIŃSKI, J., & KOTOWSKA I. E. (2015) The research method. Social Diagnosis, 2015, The Objective and Subjective Quality of Life in Poland. *Contemporary Economics*, 9(4), pp. 24-33.

[7] ZIXI, H. (2021) Poverty Prediction Through Machine Learning. Proceedings of the*2nd International Conference on E-Commerce and Internet Technology (ECIT),* pp. 314-324. IEEE.

[8] ALSHARKAWI, A., AL-FETYANI, M., DAWAS, M., SAADEH, H., & ALYAMAN, M. (2021) Poverty Classification Using Machine Learning: The Case of Jordan. *Sustainability,*13(3), 1412.

[9] LIU, M., HU, S., GE, Y., HEUVELINK, G. B., REN, Z., & HUANG, X. (2021) Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. *Spatial Statistics*, *42*, 100461.

[10] MAJEED, M. T., & MALIK, M. N. (2015). Determinants of Household Poverty: Empirical evidence from Pakistan. *The Pakistan Development Review*, 54(4I-II), pp. 701–718.

[11] FERNÁNDEZ, A., GARCIA, S., HERRERA, F., & CHAWLA, N. V. (2018) SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863-905.

[12] DRAKOS, G. (2020) *Decision tree Regressor explained in depth.* [Online]. Available from https://gdcoder.com/decision-tree-regressor-explained-in-depth/. [Accessed 05/02/2020].

[13] *Costa Rican Household Poverty Level Prediction*. (2018) [Online]. Available from https://www.kaggle.com/c/costa-rican-household-poverty-prediction

[14] CHOUBEY, V. (2020) *How to evaluate the performance of a machine learning model*. [Online]. Available from https://vijay-choubey.medium.com/how-to-evaluate-the-performance-of-a-machine-learning-model-d12ce920c365. [Accessed 25/04/2020].

[15] KUSRA, M.B., & RUDNICKI, W.R. (2010) Feature Selection with the Boruta Package[J]. *Journal of Statistical Software*, 36(11), pp. 1–13

[16]. Alex M. Goh and Xiaoyu L. Yann, (2021), "A Novel Sentiments Analysis Model Using Perceptron Classifier" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 4, pp. 01-10, DOI 10.30696/IJEEA.IX.IV.2021.01-10

[17] Dolly Daga, Haribrat Saikia, Sandipan Bhattacharjee and Bhaskar Saha, (2021), "A Conceptual Design Approach For Women Safety Through Better Communication Design" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 01-11, DOI 10.30696/IJEEA.IX.III.2021.01-11

[18] Alex M. Goh and Xiaoyu L. Yann, (2021), "Food-image Classification Using Neural Network Model" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 12-22, DOI 10.30696/IJEEA.IX.III.2021.12-22

[19] Jeevan Kumar, Rajesh Kumar Tiwari and Vijay Pandey, (2021), "Blood Sugar Detection Using Different Machine Learning Techniques" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 23-33, DOI 10.30696/IJEEA.IX.III.2021.23-33

[20] Nisarg Gupta, Prachi Deshpande, Jefferson Diaz, Siddharth Jangam, and Archana Shirke, (2021), " F-alert: Early Fire Detection Using Machine Learning Techniques" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 34-43, DOI 10.30696/IJEEA.IX.III.2021.34-43

[21]. Reeta Kumari, Dr. Ashish Kumar Sinha and Dr. Mahua Banerjee, (2021), "A Comparative Study Of Software Product Lines And Dynamic Software Product Lines" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 01-10, DOI 10.30696/IJEEA.IX.I.2021.01-10

[22]. MING AI and HAIQING LIU, (2021), "Privacy-preserving Of Electricity Data Based On Group Signature And Homomorphic Encryption" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 11-20, DOI 10.30696/IJEEA.IX.I.2021.11-20

[23]. Osman Goni, (2021), “Implementation of Local Area Network (lan) And Build A Secure Lan System For Atomic Energy Research Establishment (AERE)” Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 21-33, DOI 10.30696/IJEEA.IX.I.2021.21-33.