IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Poverty Prediction Using Machine Learning

Ajay Sharma[1*] , Jatin Rathod[2] , Rushikesh Pol[3] , Swati Gajbhiye [4]

[1]Department of Computer Science, National College, Delhi, India
[2] Department of Science and Technology, Delhi University, Delhi, India
[3] Department of Computational Sciences and Technology, Delhi University, Delhi, India

[*]*Corresponding Author:* ajayshrm91100@gmail.com *Tel.: +91 9892866246*

DOI: https://doi.org/10.26438/ijcse/v7i3.946949 | Available online at: www.ijcseonline.org

Accepted: 07/Mar/2019, Published: 31/Mar/2019

Abstract— Poverty is a classic problem in every region. It is rooted in various causes like corruption, lack of education, political instability, geographical characteristics. The success of a region is strongly influenced how big this poverty can be overcome. So that poverty reduction becomes a priority for both central and local government. There are also multiple ways to do away with it, various programs and policies began to be formulated to reduce and minimize the problem. It is extremely difficult for social programs such as this to gauge the right amount of aid that needs to be given to the right people. This problem is made exponentially more difficult when that program is dealing with the least fortunate portion of the population. This is because they cannot provide the necessary details of their income, asset or expense records to justify that they need the aid to qualify. Hence, this paper's defining question is: how to determine a method to effectively gauge the right amount of aid to be given to each household given the multitude of variables present in the vast dataset? In our work we will use supervised machine learning algorithms to a dataset to train a model which will predict the poverty based on the household level.

Keywords- Machine Learning, Random Forest, Supervised Learning, XGBoost

I. INTRODUCTION

In today's world no one wants to live poverty, yet millions of people worldwide live near poverty line. Though cities are growing faster, so there are lesser concerns in urban area, but most of the poor populations of world remain in rural area which can directly or indirectly affect social, political, economic, educational and technological processes

Poverty is a phenomenon which concerns, to a greater or lesser extent, individuals, families or households in all parts of the world. The actions of governments and social organizations are focused on helping the poor, especially households. There is a group of households living near the poverty line. Their incomes are little higher than the poverty line and these households usually do not receive any help. Actions should be focused on this group of households and to prevent their entry into poverty. The aim of this project is to analyses poverty level.

Poverty is generally associated with deprivation of basic social services such as food, health, education, knowledge, influences over one's environment and economic infrastructure that help to improve quality of life at various levels of income. – among other thing that make a difference between truly living and merely surviving. Amongst these, education is most important as it enables poor to get out of poverty over time. Other indicator that contribute to improved living standards and life expectancy are health care and environment.

In this project basically, we collected data from Kaggle source, Data is provided on the individual level where dataset consist of observation of each member in the household, so we have created household data from the individual, and apply that features to machine learning model.

The rest of the paper is organized as follows. Section II explains the related work; Section III explains the methodology; Section IV discusses the implementation details; Section V discusses the experimental results and Section VI concludes the paper.

II. RELATED WORK

The random forest algorithm is applied to customer credit evaluation, and the accuracy of the model is improved by optimizing the weight of random forest leaf nodes. At the same time, compared with the decision tree and support vector machine, the simulation results show that the improved model is effective. In the project by improving the classification of random forest, it can locate the user credit

rating more accurately than the decision tree and support vector machine algorithm [2]. The accuracy of the model is ensured by the timing feedback data to validate the model. Results not only bring benefits for power supply enterprises, but also associate with electricity, subsidies such as energy consumption by analyzing the power information of enterprises and resident in different areas, which can better understand and predict the regional and industry development situation, energy-using status and implementation effect of various policies, to provide the basis for government and industry adjustment, economic regulation.

Import the quantified data into the improved random forest, decision tree and support vector machine (SVM) model, and do 10 simulations, then use the external data obtained from formula to verify [2]. The accuracy of each training g is shown in below table. Find their average accuracy, by comparison, the accuracy of the improved random forest is better than decision tree algorithm and support vector machine algorithm. The accuracy of the more random forest algorithm is higher. But because of the random sampling of the algorithm the stability is less than the support vector machine.

Tree boosting is a highly effective and widely used machine learning method that has been shown to give state of-the-art results on many standard classification problems. [1] And the popular algorithm for tree boosting is XGBoost, this is algorithm was used for text mining project which are deciding factors for a clinician's decision towards sending a patient for an MRI scan. Here this algorithm works consistently very well than the other used algorithm has got ROC values around >0.87 across different condition and hence its proved that it is more robust in different scenario and higher dimensional dataset. [1]

However, its performance increases to a level comparable to XGB when sparse terms removal decreases from 99.9% to 99%. This is due to significant reduction of the number of variables in the bag-of-words model from thousands to a few hundred variables which is sufficient for the algorithm to perform its classification but comes at the cost of information loss as many variables in the model are dropped.

III. METHODOLOGY

Data mining is a complex analytical process of finding hidden patterns in unstructured datasets and relations between various variables and validate them to find new useful data. The main goal of data mining is data prediction and it is also one of the business application of data mining. The process of Data mining includes three phase which are as follows: (1) Data pre-processing (2) Machine learning Model Implementation (3) Using Developed model to predict data

In this paper, we will walk through a complete machine learning solution: first, get introduced to the problem, then perform a thorough data exploration analysis of the dataset, work on feature extraction, try out machine learning models, and finally, inspect the outputs of the model and draw conclusions.

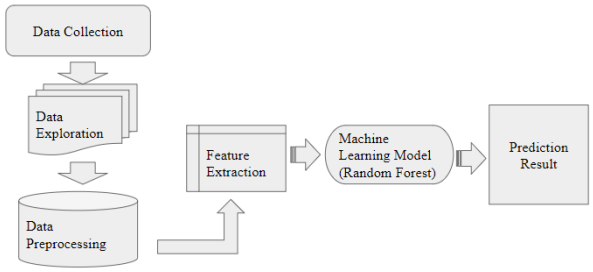

Fig 1 Work flow

From Figure-1 Work flow consist of following steps:

i. We collected data of individuals and household basis from Kaggle.
ii. After data collection, we did data exploration and data preprocessing which consist of understand the data set normalization, identify missing values etc.
iii. Later, we performed aggregation on individual data and household data.
iv. Then, we apply the Random Forest Classifier to train a model.

The aim of supervised, machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. Supervised learning algorithm has classification and regression algorithm. We are going to use classification algorithm.

Dataset Description
{train | test}.csv - The training data set and testing data set
Main dataset is divided into 2-part train.csv (with target column) and test.csv (without target column), One row represents one person in dataset in training dataset. Multiple person can be part of same household. Only prediction of head of the house is used to score the poverty level of the household [3]

This is a supervised multi-class Text classification machine learning problem:
Supervised: data provided with the labels for the train and test data
Multi-class classification: Labels are values with 4 classes

Random Forest algorithm is a supervised classification algorithm. As name suggest, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

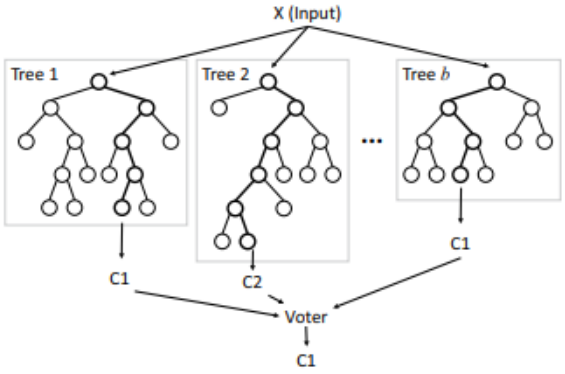

Fig 2 Random Forest

Once the data cleaning and feature engineering is done then the train data set is given to the random forest classifier then random forest algorithm select random data and random features and create decision tree. Random forest algorithm creates multiple decision tree and at the time of creating every decision tree it selects random data and random feature. This decision tree predict class according to their rules. In this classification each tree votes and the most popular class is returned.

Extreme Gradient Boosting is similar to gradient boosting framework but more efficient it has both linear model solver and tree learning solver, so what makes it fast is its capacity to parallel computation on single machine. This makes xgboost 10 times faster than existing gradient boosting is boosting kind of algorithm, it supports various objective function including regression, classification and ranking, we are going to use this algorithm in text classification problem.

Xgboost is the boosting kind of algorithm, the whole training dataset is put in to a decision tree, Once it is put into a decision tree then it creates a first weak type of classifier then for the sample which it has wrongly predicted it again passes to an another decision tree that decision tree again assigns it as a weighted sample and that is basically a second weak classifier similarly all the wrong prediction will be passed to different kind of decision trees, so this is actually happening sequentially will have multiple decision tree only the wrong prediction will be passed to a new decision tree so that it can predict it properly, as we go from one decision tree to another decision tree in a sequential way those decision trees are basically flawless weak classifier(random guessing), As shown in fig we have first weak classifier then second weak classifier and then third and then finally we will be having sequential set of weak classifier and which will finally get combined into a final classifier which will be a strong classifier.

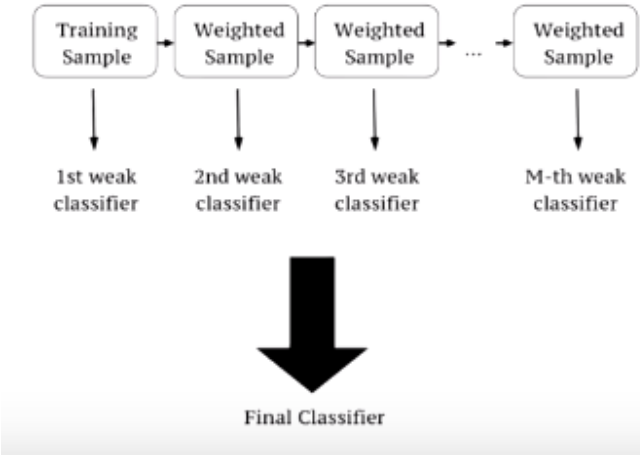

Fig 3 XG Boost

When we make a model, we'll train on a household basis with the label for each household the poverty level of the head of the household the raw data contains a mix of both household and individual characteristics and for the individual data, we will have to find a way to aggregate this for each household.

IV. RESULTS AND DISCUSSION

In this paper we have worked on random forest and xgboost algorithm to predict poverty class on head of the household level, we have come to know that Xgboost does not require feature scaling because feature scaling automatically happens inside those particular algorithms whereas random forest needs to scale feature manually. Both random forest and xgboost are robust to outliers in different scenario.

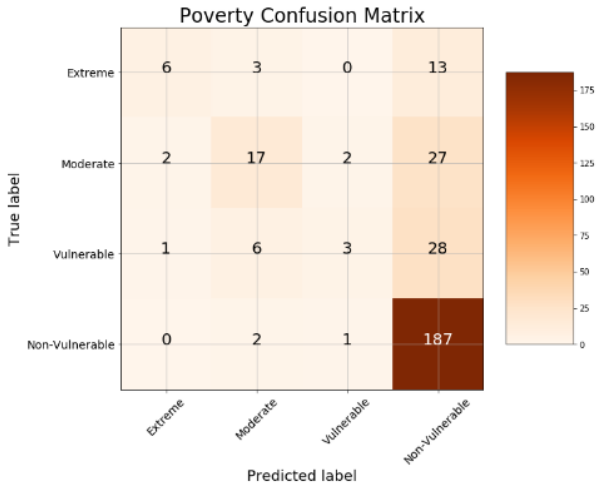

Fig 4 Xgboost Confusion Matrix

In this Experiment we observed that for random forest model we have got 0.33 macro f1 score. Then we thought that this is pretty low and to get better performance we implemented Xgboost model and we got 0.38macro f1 score.
So, we can say that xgboost classification model is more accurate than the random forest for qualitative and quantitative predictor. Here in our project we have qualitative predictor that is classification.

V. Conclusion and Future Scope

In this study we went through a step-by-step implementation of an entire data science solution to a real-world problem. Machine learning is really just a series of steps, each simple by themselves, with the overall result often extremely powerful. Our final model does well but not but overall is not extremely accurate. There might be ways to improve performance, but overall, we might not have enough data to achieve exceptional metrics. That's an important point to remember: at the end of the day, the success or failure of a data science project will rest on the quality and quantity of available data.

Although we tried only two classifier algorithms, there are still some methods that we did not implement but might prove useful

Acknowledgment

we would like to take this opportunity and express our sincere gratitude towards prof. Swati Gajbhiye maam for her guidance when required. we appreciate her valuable suggestions and support. we are also very grateful to prof. swati nadkarni, head of information technology engineering department, shah & anchor kutchhi engineering college, mumbai for her tremendous support and guidance.

References

[1]. Alwin Yaoxian Zhanga, Sean Shao Wei Lamb,c, Nan Liub,c,Yan Panga Ling Ling Chanc,d, Phua Hwee Tangc,e Development of a Radiology Decision Support System for the Classification of MRI Brain Scans IEEE 2018 Conference
[2]. Zhao, Yandong, and Xiao Ma. Study on Credit Evaluation of Electricity Users Based on Random Forest 2017 Chinese Automation Congress (CAC), 2017, doi:10.1109/cac.2017.8243614.
[3]. www.kaggle.com/c/costa-rican-household-poverty-prediction.
[4]. Amanpreet Singh, and Nanita Thakur.A review of Supervised Machine Learning Algorithm 2018 IEEE Conference.

