IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

Determining the probability of poverty levels of the Indigenous Americans and Black Americans in US using Multiple Regression

Saikat Sundar Pal[1], Soumyadeep Paul[2], Rajdeep Dey[3], Sulekha Das[4], Avijit Kumar Chaudhuri[5]

[1,2,3]UG- Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata

[4]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata

[5]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur, Habra, Kolkata

Abstract: Poverty and unequal distribution of wealth is a monumental issue that still awaits a proper solution. Poverty is prevalent all over the world. If we talk about the US, one of the most developed countries in the world, we again find poverty. The ones mostly subjected to poverty are the ethnic group of African Americans and the Native Americans. According to the 2020 census, in 10 states of U.S[1] where the majority of the African American population are found, 19.5 percent of African Americans living in the United States were living below poverty level, Native Americans have the highest poverty rate in the U.S, with one in four people living below the poverty level [2].
This Article would thus chronicle the cause behind the penury of the African Americans and the Native Americans. The percentage of people living in penury has been highlighted here. The origin of the extreme poverty levels depends upon their literacy, violent crimes, self-employed income, and community population. Data has been analyzed through Multiple Regression Analysis(MRA). The proposed model is tested on the "Communities and Crime Data Set" from the UCI Machine Learning Repository: which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime . We evaluate the model using 50–50%, 66–34% train-test splits and 10-fold cross-validation.

INTRODUCTION:

Statistics is a type of science which we need in every kind of field in our daily life i.e. also social science. We can analyze any kind of data using statistical analysis. Regression is one of the most useful statistical methods. There are six types of linear regression analyses which are simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression and discriminant analysis. Multiple linear regression uses two or more independent variables to predict something. The main objective of this paper is to select a suitable model and reasons behind the poverty level and predict how the poverty level can vary for those reasons. Post selecting the control and response variables, creating the formula and dividing the dataset into training and testing sets, creating the confusion matrix, different tests were performed and the results were noted down. All are done in the 2/3, 1/2 and 10-fold cross-validation system.

In the society of inequality and discrimination, poverty is the headache for all the growing countries. Poverty is the one of the main obstacles faced by the society in the path of development and that's why the government of every country facing it have been scrutinizing to find out it's root causes so that they could ameliorate the condition. This paper is the detailed description about the poverty level i.e. how the poverty level depends on some variable factors and how it varies depending on them. There are also some more fields on which the poverty level can depend but in this paper, mainly 4 fields are described.

Researchers have widely used machine learning (ML) methods and MLR method to predict poverty level [3]. Studies show a high accuracy of prediction of poverty data analysis using MLR methods. However, the researchers have got most precise results about the accuracy.

The database of poverty analysis has many data items [4], starting from state to violent crimes. Researchers substantiated that prediction improves with the choice of right features. Therefore, there is a need for choosing a subset of many features that best suits the task i.e. also called as trained data and test data. Among some of the several methods to select features, researchers use the MLR method because of the linear relationship between all the variables. And the result shows that the method is very much effective.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

LITERATURE REVIEW:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use data as input to predict new output values.

Intan Martina Md Ghania and Sabri Ahmadb[5] used Multiple Regression Method to Forecast Fish Landing in their research paper. Using ML method, they got two models and between those models, model 2 is better. They also got the finest result.

H. C. Hamaker[6] described a brief description about the multiple regression. How multiple linear regression works and its formulas and also where it can be used.

Julie Barber and Simon Thompson[7] depicted the usefulness of generalised linear models (GLMs) for regression analyses of cost data and they use the ML method also.

Mr. M. S. BARTLETT [8] gave a detailed information about multiple linear regression in his paper named 'FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION'.

J.T. Lin, D. Bhattacharyya and V. Kecman [9] used multiple linear regression in the field of composites machining. They are very much successful in that case.

After studying some more papers about the multiple linear regression, researchers have got a clear idea about the MLR method. That's why in this paper MLR method is used and full research paper is depending on this method.

METHODOLOGY:

- DATA: In this paper, data were taken from the UCI Machine Learning Repository. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR, https://archive.ics.uci.edu/ml/datasets/communities+and+crime. In the dataset there are total 128 columns. Out of this 5 columns are taken. 5 columns are – "PctPopUnderPov" i.e. percentage of people under the poverty level (numeric - decimal), "pctWFarmSelf" i.e. percentage of households with farm or self-employment income in 1989 (numeric - decimal), "population" i.e. population for community: (numeric - decimal), "PctBSorMore" i.e. percentage of people 25 and over with a bachelor's degree or higher education (numeric - decimal) and "ViolentCrimesPerPop" i.e. total number of violent crimes per 100K population (numeric - decimal) GOAL attribute (to be predicted).
"pctWFarmSelf" gives us a more depict answer about the people who are doing farming and living their lives. "population" tells us about the total population in a particular area per the total population in all over the area. "PctBSorMore" depicts us about the educated people who acquired bachelor's degree or higher education. "ViolentCrimesPerPop" depicts that how many criminals are there in an area per 100k population. The main motive is to find how the poverty level varies depending on the fields. Now one can say, in a growing country if there are most of the people are farmers then the growth of the country will be slow because of GDP. But one can say that farming is also a great field and it has a very much contribution in the growth of GDP but deep inside only farming can't be the main factor. That's why it may play a role in the poverty index. Population is also one of the reasons for poverty because if population will be very much high then the poverty level will increase automatically. Like these, education and crime is also related to the poverty level.

Table 1-For attributes mean and standard deviation

ATTRIBUTES	DEFINITION	MEAN	STANDARD DEVIATION
PctPopUnderPov	percentage of people under the poverty level	0.30	0.22
pctWFarmSelf	percentage of households with farm or self-employment income in 1989	0.29	0.20
population	population for community	0.05	0.12
PctBSorMore	percentage of people 25 and over with a bachelor's degree or higher education	0.36	0.20
ViolentCrimesPerPop	total number of violent crimes per 100K population	0.23	0.23

- RESEARCH METHOD: Multiple linear regression (MLR) is a type of statistical regression methods [10]. It is used to analyse the relationship between dependent variable and two or more independent variables. This method was taken for this paper because of there are more than two independent controlled variables. In this research paper, the dependent variable(y) is PctPopUnderPov while first independent variable is pctWFarmSelf(x1), second independent

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

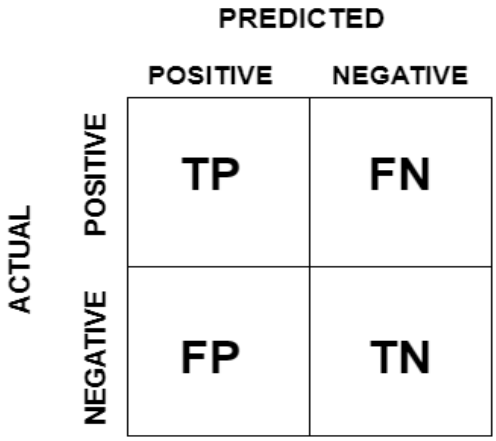
variable is population(x2), third independent variable is PctBSorMore(x3) and the last independent variable is ViolentCrimesPerPop(x4).

STEP 1 - CHECKING ASSUMPTIONS:

- PRIMARY WORK: At first, we have selected the dependent and independent variables form the dataset. After that, we have to found a linear relationship between dependent and independent variables i.e. MLR. Post this, we have taken 2/3 and 1/2 data from the dataset as trained data and rest of the data as test data. All the trained data and test data both were taken using random module.

- CROSS VALIDATION: Cross validation is a very useful technique in MLR. Cross-validation is a method in which data resampling is done to assess the generalization ability of predictive models and to prevent overfitting [11]. Overfitting is one of the insidious problem because no one can detect it unless cross-validations are carefully implemented [12]. So cross validation is one of most vital step for this research paper or verify the results of this paper correctly. We have done cross validation by dividing the total dataset into 10 parts and create 10 sub-datasets using random module. At first we are taking first 9 sub-datasets as trained data and the last one as test data. In 2nd case we took first 8 sub-datasets and the last one also as a trained data and 9th no sub-dataset as test data and so on [13].

- CONFUSION MATRIX: Confusion matrix i.e. also called as error matrix, is a type of matrix or a table where we put the results or the performance of the MLR model i.e. the test data [14]. Confusion matrix is the shortest way to see and understand the result of the model. In confusion matrix there are total four variables as – TP, TN, FP, FN. TP stands for 'true positive' which shows the number of positive data classified accurately. TN stands for 'true negative' which shows the number of negative data classified accurately. FP stands for 'false positive' which indicates the actual value is negative but predicted as positive. FP is also called as TYPE 1 ERROR. FN stands for 'false negative' which indicates the actual value is positive but predicted as negative. FN is also called as TYPE 2 ERROR [15].



- **Accuracy:** In any model, it represents the ratio of number of times the model is able to make the correct prediction to the total number of predictions.
- **Sensitivity:** It is defined as the ratio of number of times a model is able to make the positive prediction to the total number of correct predictions. In our model, it is the number of times it has made the prediction that the value of response variable(PctPopUnderPov) will be less than equals 0.2 to the number of times it has made the assumption that the value will be less than equals 0.4.
- **Specificity:** It is defined as the ratio of number of times a model is able predict that the result will be negative to the total number of times it has made the correct prediction. In this case, it is the total number of time the model is able to figure out the value of response variable(PctPopUnderPov) will be greater than 0.2 and less than equals 0.4 to the total number of correct predictions i.e. less than equals 0.4.
- **Precision:** Precision is the method in which way one can say how correctly predicted cases actually turned positive.
- **F1_SCORE:** F1 score is the measurement of accuracy and it is the harmonic mean of precision and recall. Its maximum value can be 1 and minimum value can be 0.

STEP 2 - Selecting the suitable model of multiple linear regression: MLR is used to predict something or finding the relationship between the dependent and independent variables. So in this model we are doing 2/3 and ½ data checking and also cross validation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

STEP 3 - DEVELOPING EQUATION OF MLR:

In this model, the hypothesis that is used:
H_0: None of the controlled variable X1, X2 and X3 is significantly related to Y
H_a: At least one of the controlled variable X1, X2 and X3 is significantly related to Y
The model of multiple linear regression can be represented as: [9]

$$y = a + b_1x_1 + b_2x_2 + \cdots\cdots + b_nx_n$$

$$a = \frac{\sum y^* \sum x^2 - \sum x^* \sum (x^*y)}{n\sum x^2 - (\sum x)^2}$$

$$b_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{where } i = 1,2,3\ldots\ldots n$$

Where
y = PctPopUnderPov = percentage of people under the poverty level (numeric - decimal)
a = constant variable
b_1 = coefficient of first controlled variable
b_2 = coefficient of second controlled variable
b_3 = coefficient of third controlled variable
b_4 = coefficient of fourth controlled variable and so on
x_1 = pctWFarmSelf = percentage of households with farm or self-employment income in 1989 (numeric - decimal)
x_2 = population = population for community: (numeric - decimal)
x_3 = PctBSorMore = percentage of people 25 and over with a bachelor's degree or higher education (numeric - decimal)
x_4 = ViolentCrimesPerPop = total number of violent crimes per 100K population (numeric - decimal) GOAL attribute (to be predicted)

DEVELOPING EQUATION OF CONFUSION MATRIX:

Let's take-
TP= TRUE POSITIVE
TN= TRUE NEGATIVE
FP= FALSE POSITIVE
FN= FALSE NEGATIVE
Now,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

$$F1\_Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

$$RMSE=\sqrt{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y})^2}$$

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$$

where ,

$\hat{y}$ = predicted value of y

$\bar{y}$ = mean value of y

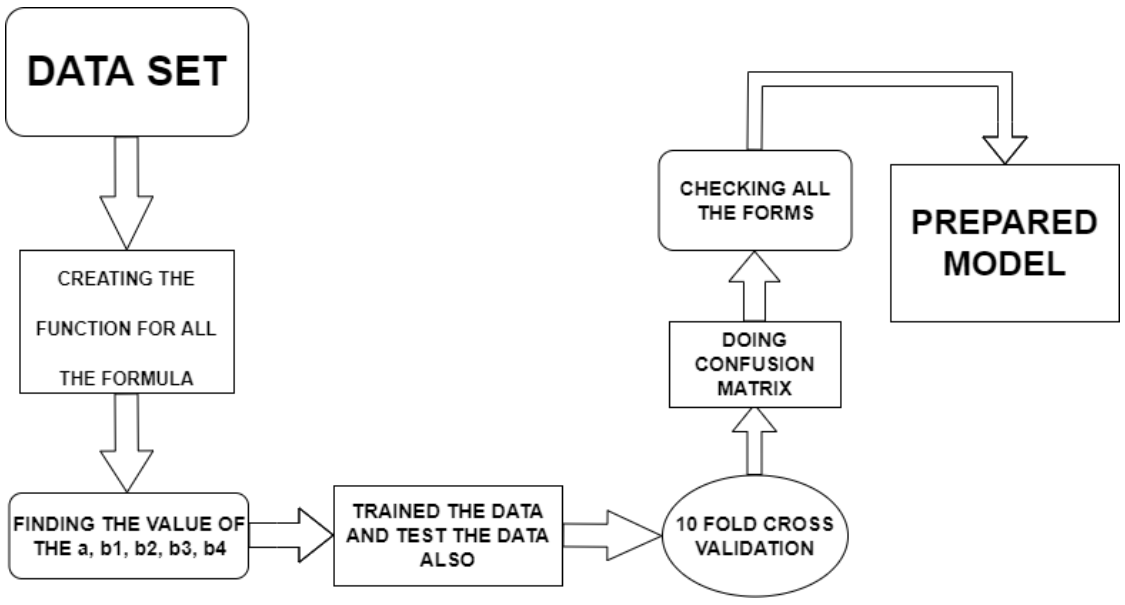
$$\text{Kappa Test} = \frac{\text{Observed Agreement - Expected Agreement}}{\text{100 - Expected Agreement}}$$

where,

Observed Agreement = %(Overall Accuracy)

$$\text{Expected Agreement} = \frac{(TP+FP) * (TP+FN) + (FN+TN) * (FP+TN)}{100}$$

FLOW CHART:



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11340

RESULTS AND DISCUSSION:

After analysing this model we get the results that are given below.

Table 2-For 66–34% train-test splits

ATTRIBUTES	VALUE RANGE(%)
Confusion Matrix	547 11 18 89
Accuracy	95.63
Sensitivity	98.02
Specificity	83.17
Standard Deviation	0.18
R_Square	0.28
Precision	96.81
F1_Score	0.97
Kappa Test	1.0

Table 3-For 50–50% train-test splits

ATTRIBUTES	VALUE RANGE(%)
Confusion Matrix	836 17 30 114
Accuracy	95.28
Sensitivity	98.01
Specificity	79.16
Standard Deviation	0.18
R_Square	0.32
Precision	96.54
F1_Score	0.97
Kappa Test	1.0

Table 4-For 10-fold cross-validation

TEST CASES	ACCURACY	SENSITIVITY	SPECIFICITY	STANDARD DEVIATION	R_SQUARE	PRECISION	F1_SCORE
01	91.63	100.0	83.34	0.08	0.53	85.59	0.92
02	87.43	97.84	78.30	0.08	0.57	79.82	0.88
03	91.45	98.01	84.69	0.07	0.35	86.84	0.92
04	89.94	96.42	85.21	0.08	0.46	82.65	0.89
05	85.92	97.75	76.36	0.08	0.47	76.99	0.86
06	90.45	95.69	85.84	0.07	0.34	85.57	0.9
07	85.92	96.77	76.41	0.08	0.67	78.26	0.87
08	89.44	100.0	80.37	0.07	0.75	81.41	0.9
09	88.94	98.79	81.89	0.08	0.53	79.61	0.88
10	88.94	98.87	80.90	0.07	0.46	80.73	0.89

CONCLUSIONS:

This paper using multiple linear regressions (MLR) to predict the poverty level. We have collected the data from UCI Machine Learning Repository based on that we made a relationship between the dependent variable and the independent

DOI: 10.17148/IJARCCE.2022.11340

variable after that we perform cross validation for more accuracy. After checking the cross validation, we move to the Confusion matrix where we compare the actual target values with those predicted by the machine learning model. Using these model, we predict the accuracy as well as sensitivity and specificity for ½, ⅔ set of data and also 10-fold cross validation. [10]

This shows that the poverty of a region can be determined by: -

- percentage of households with farm or self-employment income
- population for community
- percentage of people 25 and over with a bachelor's degree or higher education
- total number of violent crimes per 100K population

As per researchers, poverty is a very important factor for all the developing countries and also a global threat. This means that if the government pays attention to the factors like percentage of population which is still unemployed or the rate of increase of population of a community, or percentage of population that have opted for higher education and the crime rate of a particular region and tried to find out the root causes of these problems and ameliorate the conditions then the government can significantly change the poverty rate of a particular region. So if this model can predict the corrected reasons of poverty level and the results of the poverty index of a country, then any country will be able to minimize the poverty level and maximize their country's GDP i.e. growth index. So this research paper will be able to help any government to analyse their country's poverty level and then they can take decisions accordingly that can solve these severe problem.

REFERENCE:

[1] https://www.census.gov/newsroom/releases/archives/census_2000/cb01cn176.html
[2] Krogstad, J. M. (2014). One-in-four Native Americans and Alaska Natives are living inpoverty. Pew Research Center, 1-8.
[3] Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. International Journal of Intelligent Systems and Applications (IJISA), 13(4), 24-37.
[4] Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. International Journal of Intelligent Systems and Applications (IJISA), 13(4), 24-37.
[5] Ghani, I. M. M., & Ahmad, S. (2010). Stepwise multiple regression method to forecast fish landing. Procedia-Social and Behavioral Sciences, 8, 549-554.
[6] Hamaker, H. C. (1962). On multiple regression analysis. Statistica Neerlandica, 16(1), 31-56.
[7] Barber, J., & Thompson, S. (2004). Multiple regression of cost data: use of generalised linear models. Journal of health services research & policy, 9(4), 197-204.
[8] Bartlett, M. S. (1938, January). Further aspects of the theory of multiple regression. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 34, No. 1, pp. 33-40). Cambridge University Press.
[9] Lin, J. T., Bhattacharyya, D., & Kecman, V. (2003). Multiple regression and neural networks analyses in composites machining. Composites Science and Technology, 63(3-4), 539-548.
[10] Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 100011.
[11] Berrar, D. (2019). Cross-Validation.
[12] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.
[13] Stone, M. (1978). Cross-validation: A review. Statistics: A Journal of Theoretical and Applied Statistics, 9(1), 127-139.
[14] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710, 120-127.
[15] Yerpude, P. (2020). Predictive modelling of crime data set using data mining. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.
[16] Cho, Y. H. (1972). A multiple regression model for the measurement of the public policy impact on big city crime. Policy Sciences, 3(4), 435-455.
[17] Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. International Journal of Intelligent Systems and Applications (IJISA), 13(4), 24-37.
[18] Zaidi, N. A. S., Mustapha, A., Mostafa, S. A., & Razali, M. N. (2019, September). A classification approach for crime prediction. In International Conference on Applied Computing to Support Industry: Innovation and Technology (pp. 68-78). Springer, Cham.

DOI: 10.17148/IJARCCE.2022.11340

[19] Tamilarasi, P., & Rani, R. U. (2020, March). Diagnosis of crime rate against women using k-fold cross validation through machine learning. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1034-1038). IEEE.

[20] Babakura, A., Sulaiman, M. N., & Yusuf, M. A. (2014, August). Improved method of classification algorithms for crime prediction. In 2014 International Symposium on Biometrics and Security Technologies (ISBAST) (pp. 250-255). IEEE.

[21] Addy, M., Chaudhuri, A. K., & Das, A. (2020, March). Role of Data Mining techniques and MCDM model in detection and severity monitoring to serve as precautionary methodologies against 'Dengue'. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 16). IEEE.