Word Count: 3579

Plagiarism Percentage 8%

Matches

1 **World Wide Web Match**
View Link

2 **World Wide Web Match**
View Link

3 **World Wide Web Match**
View Link

4 **World Wide Web Match**
View Link

5 **World Wide Web Match**
View Link

6 **World Wide Web Match**
View Link

7 **World Wide Web Match**
View Link

8 **World Wide Web Match**
View Link

9 **World Wide Web Match**
View Link

10 **World Wide Web Match**
View Link

11 **World Wide Web Match**

View Link

Suspected Content

ESTIMATION OF POVERTY RATES IN NEW YORK CITY Harshith K Reddy

Computer Science Department (of Affiliation) PES University Bengaluru India reddyharshith90 @gmail.com Arpit Nigam Computer Science Department (of Affiliation) PES University Bengaluru India nigamarpit7000 @gmail.com Sumedh Ravi Computer Science Department (of Affiliation) PES University Bengaluru India

3

sumedhravi19@gmail.com Abstract—Poverty is an important indicator of economic well-being. Local and State policymakers use data on poverty levels to determine current economic conditions and variations based on demography of different regions. Poverty rates can be estimated as percentages or categorized into various degrees. Three models targeting point estimation, multiple classification and binary classification are put forward and compared. I. INTRODUCTION Poverty rate is defined as the proportion of people with household income less than the income threshold set for the particular year. This research paper intends to estimate the trends in poverty rates using sample data provided by the ACS estimates of census data in the city of New York. The American Community Survey (ACS) collects census data and divides it into different datasets which describe various characteristics of the census tracts. The data is collected over time periods of 1 year and 5 years. The latter encapsulates all census tracts of the state and provides highest precision by taking the largest possible sample sizes for each census tract. Therefore, we will be using data from ACS 5 year estimates for economic characteristics. This model is being built for use by various small neighbourhood communities and NGOs who would like to make an estimate of poverty rates in census tracts of interest. It is generally a difficult task to obtain data on income status of households and people in the labor force and even if it is collected, the accuracy of such data would be questionable, since a considerable proportion of people would be unwilling to disclose accurate information regarding their income sources or might not be able to provide an accurate estimate of data related to their income to non-governmental bodies. Instead of using monetary features, socio-economic factors and demographic factors such as ethnicity, means of commute, type of work, employment status and insurance coverage are much better suited for small communities to work with. II. ABBREVIATIONS AND ACRONYMS ACS -American Community Survey. It is the dataset used in estimating poverty.NGO (Non-governmental organisation). GLM and MLR stand for generalised linear model and Multiple Linear Regression respectively.β and wald are the regression coefficients and test statistic respectively. PCA (Principal Component Analysis). RFR (random forest regression), SVM (support vector machine). III. LITERATURE REVIEW Citation:

E. Xhafaj and I. Nurja,

11

“DETERMINATION OF THE KEY FACTORS THAT INFLUENCE POVERTY THROUGH ECONOMETRIC MODELS”, ESJ, vol. 10, no. 24, Aug. 2014. Abstract This paper tries to see the foremost necessary factors which will influence economic conditions through the models

Logistic regression and Log -Linear regression. The data is obtained from the

1

Living Standards Measurement Study (LSMS) for 2008 that includes 3600 households interviewed in Albania. For the log linear model, the target variable is determined by the total expenses of the household per capita.

The predictor variables are various demographic variables like household size and education level of the family head and variables describing the geographic area as in urban and rural. The logistic model uses the binary classification of poverty, either not poor or poor as its dependent variable, with the dependent variables being the same. Results From the table given below, for the logistic regression we can infer

that household size is a statistically significant variable (β = 0.548 Wald = 41092,929 p =0.000) and its positive coefficient indicates that 1

probability of poverty increases with increasing household size. Coming to the Education level of the head of the family, we find it to be

statistically significant(B = -0.350 Wald = 6338,956 p =0.000) and 1

negatively correlated with probability of being poor. This implies households with the breadwinners having higher levels of education have lower chance of being poor. When taking female headed households into consideration, we find that such households have a lesser chance of being under poverty, concluded by looking

at the negative XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE coefficient of the 8

predictor variable Wald=560.004, p=0.000). table 1: Logistic Regression model evaluation β Wald (β=-0,381

p Zone 0.094 100.909 0.000 Education of household head -0.350 6338.956 0.000 Female head -0.381 560.004 0.000 Household size 0.548 41092.929 0.000 Constant -4.146 61346.523 0.000 1

The logistic model also shows that households

in rural areas have a higher probability of 1

falling under the poverty line. Similar results were obtained from the log linear regression model. Conclusion The two models used in this research paper, the log linear model and the logit model, try to predict poverty by finding estimates of gross family expenditure and classification as poor or not-poor respectively. For both models, we find that size of the family, education level of the family head, female headed families and the geographic zone of the household are important factors in determining the poverty levels. Shortcomings We

can predict poverty based on three different models. We can estimate poverty levels, we can make sub-classes of poverty with varying degrees of poverty levels or else we can directly classify poverty as status of being poor or not poor. Highest preference of the model would be the regression estimation model because point estimates let us know exactly what percentage of the households are under poverty and how far are they from the poverty line. But point estimates tend to have a higher margin of error associated with them. This decreases as we decrease the range of outcomes possible. In other words, classification models tend to give more accurate results when there is no direct linear relationship between the predictor and target variables. This research paper has built such a classification model by binary classification. However, binary classification holds far lesser information than point estimates. What we require is a compromise between the two- adequate amount of information with a relatively lower margin of error. We can achieve this by increasing the range of classification outputs using models like random forests and neural networks. IV. PROPOSED SOLUTION About The Dataset The dataset used in this analysis is

of selected economic characteristics (ACSDP5Y2018.DP03, U.S. Census Bureau) for the city of New York. The features of this dataset describe various economic parameters of the entire city of New York. Each row in the table represents a particular geographic region known as a 'census tract'. The boundaries of these census tracts are mapped by the Census Bureau and generally consist of populations ranging between 1000 to 5000 residents. Since it is not possible to obtain the data of the entire population in each census tract, the ACS collects data by taking a sample from the population. Hence, it is important to know beforehand that the entries of the features for each row are estimates of the original parameters. Hence, they are subject to sampling variation and have a margin of error associated with them. The ACS has adopted a standard confidence interval of 90% for defining the margin of error. Any model made with such estimates will have its accuracy decreased since the feature sets are themselves estimates of the parameters. The feature set of this dataset can be categorized broadly into 6 categories- ● ● ● Employment status Occupation Nature of work (private, government employed, self-employed, etc.) ● ● ● ● Means of commute Income and benefits Insurance coverage Poverty status Each of these principal categories has numerous features explaining them which subclassify these categories based on age, gender, and other sub-categories. To conclude the summary of the dataset, it is worth mentioning that each feature described above is represented both as a numerical estimate and a percentage estimate with their associated margin of errors also being given separate columns. This is helpful in cases where the model only requires data having low margin of error which can be filtered out from the original dataset. Data Pre-processing Pre-processing of the dataset includes missing values and outlier treatment, feature scaling, and data reduction. Missing values in this particular dataset are caused because of two reasons- due to the population in a census tract being zero, or the estimate of a feature(column) in a tract could not be produced due to a high margin of error. Such cases are marked with specific non-numerical symbols that indicate the reasons for missing values. When missing values are due to the population in the census tract being zero, we simply drop the concerned rows because we cannot do any analysis on areas that do not have residents living there. As for missing values due to the high margin of error, we can choose to ignore them if we prefer accuracy over retaining all data. Otherwise, we try and fill in the missing values using data from other features in such rows. For this paper, the second approach, imputing missing values based on other columns and rows have been used. Imputation of missing values can be done elegantly with the help of scikit learn library's Iterative Imputation class. Iterative imputation refers to a process where missing values of a feature are predicted by modeling said feature as a regressive function of other features. Each feature is

imputed sequentially, one after the other, allowing features with imputed values earlier to be used as part of a model in predicting features. As part of feature scaling, the dataset is standardized with each feature having

a mean value of 0 and a standard deviation of 5

1. Standardization helps bring features having different scales of measurement to a single scale. This helps reduce bias among features due to the varying scales, which leads to the better fitting of training data and improved accuracy. After cleaning the dataset, 117 features remain, quite a few of which are not correlated to the target variable. However, dropping a large number of features to reduce the complexity of the model may result in a loss of information. Hence, in order to reduce the size of the dataset without compromising the information provided PCA (principal component analysis) can be used. PCA transforms the existing dimensions associated with the features to a new feature subspace where the components are linear combinations of the original features. We can then select the number of components to keep based on how highly correlated they are with the target variable. However, applying PCA to the dataset resulted in loss of correlation between the dependent and independent variables. Therefore, the linear regression model can be built first and then variables can be discarded based on their t-test statistical significance. Multiple

Linear Regression model A linear regression model describes the relationship between the predictors and 5

the target variable by fitting a linear equation to the data. In multiple linear regression, the number of predictors is more than one. For a dataset containing n predictor variables, the MLR model fits an n-dimensional subspace to the data. The most common method used to fit the regression line to the dataset is the Ordinary Least Squares (OLS) fit method. The MLR model has two versions, the first one having features describing the income status of the population of census tracts, and the other without any income-related features. It would make sense to foreshadow that the model which has income-related features will be far more accurate than the model which does not have any income-related features, since income status measured as the central tendency of a census tract would be highly correlated with the poverty rate of that tract. A census tract having higher median household income would usually have lower poverty rates. Therefore, instead of asking which model is preferable for obtaining more accurate results, it would be reasonably better to ask whether it would be worth removing income predictors for the sake of reduced model complexity and increased convenience while sampling data from people without considerably sacrificing model accuracy. This is an important question to consider since NGO’s and small communities would generally prefer to collect non-income related data. Thus, it is crucial to know beforehand whether the model performs well enough when compared to a model based on income predictors. Logistic Regression model Logistic Regression, additionally called Logit Regression, could be a mathematical model that's employed

in statistics to estimate (guess) the chance of an occurrence occurring having been given some previous information. supplying Regression works with binary information, wherever either the event happens (1) or the event doesn't happen (0). thus given some feature x, it tries to search out whether or not some event y happens or not. 2

thus y will either be zero or

one. Since logistic regression was a binary classifier, we classified poverty into two levels: ● level 0: less than equal to mean in the poverty metric. ● level 1: greater than mean in the poverty metric. The regression model was trained on 80% of the dataset and the remaining was used for testing. Support Vector Machine

model Support Vector Machine (SVM) could be a supervised machine learning formula which will be used each for classification and regression challenges. 7

However, it's largely employed in classification issues. Within the SVM

formula, we tend to plot every information item as a degree in n- dimensional area (where n is the variety of options you have) with 4

every feature being worth getting selected as a coordinate. The SVM formula incorporates a technique referred to as the kernel trick, that takes low dimensional input area and transforms it into a better dimensional area Since SVM was a binary classifier similar to logistic regression, we classified poverty into two levels: ● level 0: less than equal to mean in the poverty metric. ● level 1: greater than mean in the poverty metric. The regression model was trained on 80% of the dataset and the remaining was used for testing. Random Forest model Random forest could be a supervised machine learning formula, typically it builds AN ensemble of call trees trained with the fabric methodology. The thought of the fabric methodology is that it's a mixture of learning models that will increase the results. Rather like SVM, the Random forest formula can even be used for classification and regression issues. Random forest adds extra randomness to the model whereas growing the trees. rather than looking for the foremost necessary feature whereas rending a node, it searches for the most effective feature among a random set of options. This leads to a large diversity that typically leads to a stronger model. For our approach we classified poverty into three levels: ● level 0: less than equal to 25% in the poverty metric. ● level 1: between 25 to 50% in the poverty metric. ● level 2: above 50% The Random Forest model was trained on 70% of the dataset and the remaining 30% of the dataset was used for testing, V. EXPERIMENTAL RESULTS The following table summarizes the MLR model with Income predictors p-val(F) 0.00 kurtosis 50.654 After analysing the p-values for the t-test for individual predictors, insignificant features were dropped.The model has an R-squared value of 0.821, which implies that nearly 82% of the variance in the Target variable is explained by the Linear equation. Adjusted R-squared is almost equal to the R- squared, implying that almost all of the independent variables are significant. The high value of the F-statistic and its negligibly small p-value indicates that the model as a whole is much more significant than a model that does any independent variables.

The omnibus test is a chi-square test of the 9

current model versus the intercept model. The significance value of less than 0.05 indicates that the current 6

model performs better than a model having only the intercept. The skewness and kurtosis values are also highly positive, indicating that the target variable is right-skewed and not normally distributed(fig.1). This results in the residuals deviating from a normal distribution (fig.3) and their variance depending on the value of the independent variables (heteroscedasticity) as seen in fig.2. Although the model does not follow the assumptions required for analyzing an MLR model, the high value of R-squared indicates a good fit of data given the data does not lie at the extreme ends of the distribution. fig. 1: distribution of poverty rates table 2:Multiple Linear Regression model summary(Income Predictors) fig. 2: predicted values vs. residual plot R2 0.821 omnibus 3972.948 Adj. R2 0.819 p-val(omb) 0.00 F-score 595.9 skew 3.249 fig.3 Probability plot of residuals The following table summarizes the MLR model without income predictors: table 3:Multiple Linear Regression model summary(without income predictors) R2 0.727 omnibus 2748.998 Adj. R2 0.723 p-val(omb) 0.00 F-score 205.4 skew 2.038 p-val(F) 0.00 kurtosis 26.880 The R-squared value for the model is 0.69 which is 10% lesser than the R-squared value for the MLR with income predictors. There is a considerable decline in the fit of the model, which is expected. Fitting both the models on the testing data, it is revealed that the root mean squared error is 5.55 for the model with income predictors and 6.67 for the model without the income predictors. This implies that the poverty rates which are predicted would roughly have a margin of error of 5.5 for the first model and 6.6 for the second model. The model without income predictors has an approximately 15% higher margin of error than its counterpart. For the Classifying problem, the results obtained were identical for logistic regression and SVM, the train test split of 0.2 was used. The following table summarizes the results obtained for the same without the Income predictors: table 4: Logistic regression and SVM model summary precision recall support 0 0.89 0.79 364 1 0.88 0.94 607 The model has an accuracy of 0.88 when tested on the testing data, which was 20% of the entire data set. A precision of 0.89 indicates that the model was right 89% of the times when the output belonged to class 0, similarly it was right 88% of the times when predicted class was 1. A recall of 0.79 indicates that the model was able to correctly identify 79% of all of the class 0, similarly it was able to correctly identify 94% of all class 1's. For the Random forest model the following results were obtained (table 5): table 5: RFR model summary precision recall support 0 0.94 0.99 0.96 1 0.72 0.54 0.61 2 0.50 0.05 0.09 The model has an accuracy of 0.92 when tested on the testing data, which was 30% of the entire data set. A precision of 0.94 indicates that the model was right 94% of the times when the output belonged to class 0, similarly it was right 72% of the times when predicted class was 1 and 2 when the output belonged to class 2. A recall of 0.99 indicates that the model was able to correctly identify 99% of all of the class 0, similarly it was able to correctly identify 54% of all class 1's and only 5% for class 2. The model generally failed to predict the poverty rates at the extreme classes(>50% poverty) since most of the data set consisted of class 0(less than 25% poverty rate) and class 1(less than 50% poverty), therefore we couldn't capture enough of class 2 in the training and testing data set. VI. CONCLUSION The motive of this research paper was to build a model which provides sufficient information about poverty levels of census tracts without compromising on accuracy of the estimates. Four models were built using different estimation techniques. The first model, a linear regression model provides point estimates of the poverty rates of the census tracts. The second model, a random forest classifier, classified the poverty rate in three different classes(below 25% poverty, below 50% poverty and above 50% poverty). Since the poverty rate samples have extreme right skew, the classes are divided non-uniformly to make the sample distribution more even. The third and fourth models are binary classifiers, classifying poverty as above average poverty and below average poverty, where average poverty rate was deduced from the sample data. We find that as the amount of information available from the estimate decreases(the highest amount of information available from point estimates, and lowest information gain from binary classifiers), accuracy of the model increases. We can conclude that the random forest classifier does an excellent job preserving information about the poverty status without

compromising on the accuracy of the classification. Contributions The research paper analysis and analysis of the data was done by Sumedh Ravi. Harshith Reddy was responsible for data preprocessing and building the multiple regression model. Arpit Nigam has built the three classification models (Random forest, SVM and logit regression). Acknowledgements

We would like to express special thanks to 2

our Data analytics professor, Prof. Bharathi R for guiding us throughout the research. We would also like to thank the US Government for collecting regular Census data and providing thorough explanation on the features of the data. References [1]

E. Xhafaj and I. Nurja, 11

“DETERMINATION OF THE KEY FACTORS THAT INFLUENCE POVERTY THROUGH ECONOMETRIC MODELS”, ESJ, vol. 10, no. 24, Aug. 2014. [2] Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J. M., and Ballon, P. “MULTIDIMENSIONAL POVERTY MEASUREMENT AND ANALYSIS”. Oxford: Oxford University Press, ch. 10,2015. [3] Xizhi Zhao 1,2,3, Bailang Yu 1,2,* , Yan Liu 3,* , Zuoqi Chen 1,2 , Qiaoxuan Li 1,2 , Congxiao Wang 1,2 and Jianping Wu 1,2. “ESTIMATION OF POVERTY USING RANDOM FOREST REGRESSION WITH MULTI-SOURCE DATA: A CASE STUDY IN BANGLADESH”,2019.