Analysis on Predicting Poverty using Logistic Regression

18/PCSA/102

2/25/2020

DATA CLEANING

```
library(vcd)

## Loading required package: grid

library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

library(caTools)
library(e1071)
options(warn=-1)
#install.packages("ROCR")

#IMPORT DATASET
dataset<-read.csv(file.choose(),header=TRUE)
str(dataset)

## 'data.frame':    1144 obs. of  28 variables:
##  $ row_id                          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ country                         : Factor w/ 7 levels "A","C","D","
F",..: 1 2 5 3 4 2 1 3 4 1 ...
##  $ is_urban                        : logi  FALSE FALSE TRUE FALSE FAL
SE FALSE ...
##  $ age                             : int  61 31 50 44 37 17 50 58 32
41 ...
##  $ Gender                          : Factor w/ 3 levels "female","mal
e",..: 2 1 2 1 1 1 1 2 1 2 ...
##  $ married                         : logi  TRUE TRUE FALSE FALSE FALS
E FALSE ...
##  $ education_level                 : int  0 1 2 1 2 2 1 0 2 3 ...
##  $ literacy                        : logi  FALSE TRUE TRUE TRUE TRUE
```

```
TRUE ...
##  $ employed_last_year                   : logi  FALSE TRUE TRUE FALSE TRUE
FALSE ...
##  $ employment_category_last_year        : Factor w/ 5 levels "employed","h
ousewife_or_student",..: 2 1 1 2 1 2 1 4 1 1 ...
##  $ employment_type_last_year            : Factor w/ 5 levels "irregular_se
asonal",..: 2 4 5 2 4 2 5 2 5 5 ...
##  $ share_hh_income_provided             : int  1 3 3 5 5 1 5 2 5 5 ...
##  $ num_times_borrowed_last_year         : int  1 0 1 3 1 1 2 1 0 0 ...
##  $ borrowing_recency                    : int  2 1 2 2 2 2 2 2 0 0 ...
##  $ formal_savings                       : logi  FALSE TRUE TRUE FALSE TRUE
FALSE ...
##  $ informal_savings                     : logi  FALSE FALSE FALSE FALSE FA
LSE FALSE ...
##  $ has_insurance                        : logi  FALSE FALSE TRUE FALSE TRU
E FALSE ...
##  $ has_investment                       : logi  FALSE TRUE TRUE FALSE TRUE
FALSE ...
##  $ borrowed_for_emergency_last_year     : logi  FALSE TRUE FALSE FALSE FAL
SE TRUE ...
##  $ borrowed_for_daily_expenses_last_year: logi  TRUE TRUE TRUE TRUE TRUE T
RUE ...
##  $ phone_technology                     : int  0 3 3 2 2 1 0 0 3 3 ...
##  $ phone_ownership                      : int  1 2 2 2 2 2 0 0 2 2 ...
##  $ reg_bank_acct                        : logi  FALSE TRUE TRUE FALSE TRUE
FALSE ...
##  $ active_bank_user                     : logi  FALSE FALSE TRUE FALSE TRU
E FALSE ...
##  $ num_formal_institutions_last_year    : int  0 0 3 1 1 0 0 0 0 2 ...
##  $ num_informal_institutions_last_year  : int  0 0 1 0 2 0 0 0 0 0 ...
##  $ num_financial_activities_last_year   : int  0 1 8 1 6 0 0 0 0 2 ...
##  $ Poverty                              : int  0 1 1 1 1 1 0 0 1 1 ...
```

```
dataset<-na.omit(dataset)
attach(dataset)
```

TEST OF INDEPENDENCE

```
#TEST OF INDEPENDENCE
mytable<-xtabs(~employed_last_year+literacy,data=dataset)
chisq.test(mytable)
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  mytable
## X-squared = 148.41, df = 1, p-value < 2.2e-16
```

```
fisher.test(mytable)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  mytable
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  3.926158 6.932517
## sample estimates:
## odds ratio 
##   5.205768
```

```
#HAS_INSURANCE VS LITERACY
mytab<-xtabs(~has_insurance+literacy,dataset)
chisq.test(mytab)
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  mytab
## X-squared = 143.34, df = 1, p-value < 2.2e-16
```

```
fisher.test(mytab)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  mytab
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   7.409657 20.770279
## sample estimates:
## odds ratio 
##   12.07717
```

```
#URBAN VS LITERACY
my<-xtabs(~is_urban+literacy)
chisq.test(my)
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  my
## X-squared = 55.679, df = 1, p-value = 8.535e-14
```

```
fisher.test(my)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  my
## p-value = 1.938e-14
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.181036 3.942155
## sample estimates:
## odds ratio
##   2.922139

#FORMAL SAVING VS EMPLOYMENT CATEGORY
tab<-xtabs(~formal_savings+employment_category_last_year,dataset)
chisq.test(tab)

## 
##  Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 268.55, df = 4, p-value < 2.2e-16

#HAS INVESTMENT VS EMPLOYMENT
tabl<-xtabs(~has_investment+employment_category_last_year,dataset)
chisq.test(tabl)

## 
##  Pearson's Chi-squared test
## 
## data:  tabl
## X-squared = 292.52, df = 4, p-value < 2.2e-16

assocstats(tabl)

##                     X^2 df P(> X^2)
## Likelihood Ratio 337.23  4        0
## Pearson          292.52  4        0
## 
## Phi-Coefficient   : NA 
## Contingency Coeff.: 0.456 
## Cramer's V        : 0.512

#HAS INSURANCE VS LITERACY
t<-xtabs(~has_insurance+literacy,dataset)
chisq.test(t)

## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  t
## X-squared = 143.34, df = 1, p-value < 2.2e-16

assocstats(t)

##                     X^2 df P(> X^2)
## Likelihood Ratio 175.69  1        0
## Pearson           145.04  1        0
##
## Phi-Coefficient   : 0.361
## Contingency Coeff.: 0.339
## Cramer's V        : 0.361
```

```
fisher.test(t)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   7.409657 20.770279
## sample estimates:
## odds ratio
##   12.07717
```

```
#HAS INSURANCE VS HAS INVESTMENT
ta<-xtabs(~has_insurance+has_investment,dataset)
chisq.test(ta)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ta
## X-squared = 80.139, df = 1, p-value < 2.2e-16
```

```
assocstats(ta)
```

```
##                      X^2 df P(> X^2)
## Likelihood Ratio 80.970  1        0
## Pearson          81.327  1        0
##
## Phi-Coefficient   : 0.27
## Contingency Coeff.: 0.261
## Cramer's V        : 0.27
```

```
#BORROWED FOR DAILY VS EMERGENCY
a<-xtabs(~borrowed_for_daily_expenses_last_year+borrowed_for_emergency_last_y
ear,dataset)
chisq.test(a)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  a
## X-squared = 22.98, df = 1, p-value = 1.637e-06
```

```r
#INCOME VS BORROWING RECENCY
b<-xtabs(~share_hh_income_provided+borrowing_recency,dataset)
chisq.test(b)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  b
## X-squared = 44.741, df = 8, p-value = 4.12e-07
```

```r
#EMP TYPE VS EMP CATEGORY
c<-xtabs(~employment_type_last_year+employment_category_last_year,dataset)
chisq.test(c)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  c
## X-squared = 2228, df = 16, p-value < 2.2e-16
```

```r
assocstats(c)
```

```
##                    X^2 df P(> X^2)
## Likelihood Ratio 1786 16        0
## Pearson          2228 16        0
## 
## Phi-Coefficient   : NA 
## Contingency Coeff.: 0.816 
## Cramer's V        : 0.707
```

```r
#NUM FINANCIAL ACT VS NUM FORMAL INSTITUTION
d<-xtabs(~num_financial_activities_last_year+num_formal_institutions_last_yea
r,dataset)
chisq.test(d)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  d
## X-squared = 1560.7, df = 60, p-value < 2.2e-16
```

```r
assocstats(d)
```

```
##                      X^2 df P(> X^2)
## Likelihood Ratio 1676.2 60        0
## Pearson          1560.7 60        0
## 
## Phi-Coefficient   : NA 
## Contingency Coeff.: 0.764 
## Cramer's V        : 0.483
```

```
#ACTIVE BANK USER VS LITERACY
e<-xtabs(~active_bank_user+literacy,dataset)
chisq.test(e)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  e
## X-squared = 246.81, df = 1, p-value < 2.2e-16
```

```
assocstats(e)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 298.21  1        0
## Pearson          248.89  1        0
##
## Phi-Coefficient   : 0.473
## Contingency Coeff.: 0.427
## Cramer's V        : 0.473
```

```
fisher.test(e)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  e
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  13.58107 39.03870
## sample estimates:
## odds ratio
##   22.37564
```

```
#REG BANK VS LITERACY
f<-xtabs(~reg_bank_acct+literacy,dataset)
chisq.test(f)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  f
## X-squared = 321.4, df = 1, p-value < 2.2e-16
```

```
assocstats(f)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 365.39  1        0
## Pearson          323.73  1        0
##
## Phi-Coefficient   : 0.539
## Contingency Coeff.: 0.475
## Cramer's V        : 0.539

fisher.test(f)

##
##  Fisher's Exact Test for Count Data
##
## data:  f
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  14.3586 33.9881
## sample estimates:
## odds ratio
##   21.73673

dataset$Poverty<-as.factor(dataset$Poverty)
```

IMPLEMENTATION

```
#creating training and testing data
set.seed(300)
split<-sample.split(dataset,SplitRatio = 0.80)
split

##  [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TR
UE
## [13]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TR
UE
## [25]  TRUE  TRUE  TRUE FALSE

train<-subset(dataset,split=="TRUE")
test<-subset(dataset,split=="FALSE")
table(train$Poverty)

##
##   0   1
## 268 607

table(test$Poverty)

##
##   0   1
##  76 163

#BUILT THE LOGISTIC MODEL
mymodel<-glm(Poverty~employed_last_year+education_level+informal_savings+has_
investment+phone_technology+reg_bank_acct+num_financial_activities_last_year+
is_urban, family='binomial', data=train,maxit=100)

mymodel
```

```
##
## Call:  glm(formula = Poverty ~ employed_last_year + education_level +
##     informal_savings + has_investment + phone_technology + reg_bank_acct +
##     num_financial_activities_last_year + is_urban, family = "binomial",
##     data = train, maxit = 100)
##
## Coefficients:
##                        (Intercept)              employed_last_yearTRUE
##                            -2.7696                              0.1638
##                    education_level                informal_savingsTRUE
##                             1.7819                              0.5530
##                 has_investmentTRUE                    phone_technology
##                             0.5515                              0.4044
##                  reg_bank_acctTRUE  num_financial_activities_last_year
##                             0.5861                              0.1091
##                       is_urbanTRUE
##                             0.3605
##
## Degrees of Freedom: 874 Total (i.e. Null);  866 Residual
## Null Deviance:       1078
## Residual Deviance: 455.1     AIC: 473.1
```

```
summary(mymodel)
```

```
##
## Call:
## glm(formula = Poverty ~ employed_last_year + education_level +
##     informal_savings + has_investment + phone_technology + reg_bank_acct +
##     num_financial_activities_last_year + is_urban, family = "binomial",
##     data = train, maxit = 100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6255  -0.3775   0.1035   0.2839   2.3793
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -2.76964    0.24566 -11.274  < 2e-16 **
*
## employed_last_yearTRUE              0.16376    0.29644   0.552  0.58065
## education_level                     1.78191    0.17246  10.332  < 2e-16 **
*
## informal_savingsTRUE                0.55304    0.35583   1.554  0.12013
## has_investmentTRUE                  0.55147    0.34025   1.621  0.10507
## phone_technology                    0.40440    0.13706   2.950  0.00317 **
## reg_bank_acctTRUE                   0.58612    0.41455   1.414  0.15740
## num_financial_activities_last_year  0.10908    0.09265   1.177  0.23905
## is_urbanTRUE                        0.36053    0.31178   1.156  0.24753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1078.17  on 874  degrees of freedom
## Residual deviance:  455.07  on 866  degrees of freedom
## AIC: 473.07
##
## Number of Fisher Scoring iterations: 6
```

```
#PREDICTION
restest<-predict(mymodel,test,type="response")


confmatrix<-table(Actual_Value=test$Poverty,Predicted_Value=restest>=0.5)
confmatrix
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0    68    8
##            1     9  154
```

```
#ACCURACY OF THE MODEL
accuracy<-sum(diag(confmatrix))/sum(confmatrix)
print(paste("Accuracy of the test",accuracy))
```

```
## [1] "Accuracy of the test 0.928870292887029"
```

Analysis On Escaping and Falling into Poverty
Using Logistic Regression

ABSTRACT

The study explains the analysis of the poverty dataset, what are the factors that influence the poverty rate and what are the factors that decline the poverty rate are examined through the analysis. The dataset used for the analysis has individual household data which helps in identifying the factors that lead to poverty of each individual. Descriptive statistic technique was applied to the dataset to find the mean, median, minimum value and maximum value of each variables. In order to find out the dependency between the variables in the dataset the test of independence and correlation was applied on the variables. The logistic Regression algorithm is used in predicting the poverty. Logistic Regression is appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal or ordinal independent variables. The algorithm is implemented using the R language and the accuracy produced was 92%.

Keywords: Logistic Regression, Test of Independence, Poverty, Prediction

INTRODUCTION

Many countries have remarkable achievements in dealing with the issues of poverty, as poverty reduction has been a major goal of development policy after independence. The role of structural characteristics like caste, religion and region are known on temporal understanding of poverty, the household data of each individual plays important role in understanding the poverty. This study is using the poverty dataset which helps in identifying the households those fallen into poverty. It involves 36 attributes and 8059 instance of records present in the dataset. Data cleaning is the first step carried out in the dataset, all the NA and wrong data from the dataset was removed.

The dataset comprises of poverty probability rate ranging from 0.5 to 1. Basic functions like mean, median, range, str, dim and mode has been applied on the dataset in order to understand the data. To find the relationship between the variables in the dataset correlation function and test of independence was applied on the dataset. The result showed that illiteracy and poverty probability are more dependent on each other followed by the employment type and poverty probability. The analysis showed that people who live in rural area fall under the poverty compared to the people living in urban areas. The people who are married have higher rate of poverty compared to unmarried. If people are illiterate, they are under high poverty rate compared to literate. The people who have investments, insurance, informal savings and bank account are escaping from the poverty rate. After the analysis of the data the attributes which are more dependent on each other was identified and what are the attributes that influence the poverty was identified. The study uses logistic regression algorithm for predicting the poverty.

LITERATURE REVIEW

In spite of significant reduction in the poverty rate in last five decades, India still continues to have the highest incidents of poverty. Poverty continues to be a significant issue in India, in spite of its being one of the fastest growing economies in the world. The absence of large-scale data on same households at different points of time, has deprived researchers of the deeper analysis of household dynamics in general and poverty in specific. The Indian Human Development Survey (IHDS) database provided the opportunity to get household information. The unique IHDS database provides panel data at two points of time in the year 2005-06 and 2011-12. The IHDS has two datasets –IHDS-I and IHDS-II. These two datasets are linked based on linkhh (linking household dataset) with 5 ids. The five IDs are STATEID (state code), DISTID (District code), PSUID (Primary sampling unit), HHID (Household id) and HHSPILITID (Split household id). The study explains about people who have fallen into poverty and people who have escaped from poverty in-between the years 2006 to 2012.

This study is using the poverty line defined by Tendulkar Committee report based on consumption expenditure, for both the points of time. The feature selection was done based on recent literature in last five years, 22 attributes were selected for the analysis like bank account, urban rural, caste, education, cow, insurance, buffalos, sheep and drinking water etc. the data representation was done by grouping each record into increased state, decreased state, no change and empty state. If there is increase in a particular factor from 2005 to 2012, it labelled as “inc”. For example-if number of cows has increased then the value in that particular row is labelled as “inc”. If there is a decrease in a factor from 2005 to 2012, then it is labelled as “dec”. If there is nochange in a factor from 2005 to 2012, it is labelled as “nochange”. If it is empty in both the years 2005 and 2012, it is labelled as “empty”. Data distribution and classification is done by dividing the Falling in group into two classes as the distribution of consumption expenditure is skewed to one side. Comfort class and average class is taken. Escaping group has been divided into three classes. Destitute class, the average class and edge class is taken.

As the distribution of data among classes is uneven smote technique is used to balance the data. The Relevant features were extracted using Machine Learning technique using Entropy based Info-Gain. WEKA 3.8 is an open source tool used for this study. A decision tree was constructed using random forest technique to identify which feature plays important role in the prediction of poverty. This provides the rank order of the attributes based on Information Gain for poverty analysis, out of which 10 attribute was selected for the analysis. The accuracy is calculated using 10-fold cross – validation. the F-score and Roc are above 71% and 85% respectively, it explains that the features taken for explaining both falling and escaping from poverty are relevant. From the study the features that helps to identify Falling into and escaping poverty was Buffalos, goats, cows, other animal’s, education and caste. Falling into poverty was poultry, house type, credit saving and drinking water. Escaping poverty- Rural_Urban, Toilet, Bank and LIC insurance explanatory.

ANALYSIS ON DATASET

The dataset comprised of 32 attributes and 3045 instance of record. The response variable was poverty and the predictor variables were literacy, employment type, education level, borrowed for emergency and employment category etc. The dataset consists of one response variable and multiple predictor variable for predicting the poverty. First and foremost step, applied on the dataset was data cleaning and preprocessing, removing all the unwanted records like NA and renaming the column names. Once data cleaning has been applied on the dataset, basic functions like head, tail, summary, class and standard deviation was identified. The test of independence was identified using the chi-square function, associate function and fisher function. The descriptive statistic was applied on the dataset using functions like sapply and describe function. Next data visualization is done using the functions like plots, histogram, boxplot, dot plot, bar plot, stacked bar plot, spinogram, pie chart, fan plot and density plot. Based on the analysis done on the dataset was able identify what are the variables that play significant role in leading to poverty. The dataset consists of 3075 records of people who are illiterate and 5119 number of people are literate. It contains 5616 records of rural data and 2784 records of urban data. Female who are illiterate are 1954 in number and 2493 records are literate. The number of females who are employed are 2029 in number and unemployed are 2418 in number. The number of male who are employed are 2762 in number and unemployed are 850 in number. Employment rate is more compared to unemployment. The age group from 20-60 is suffering more from poverty compared to age group from 70-80. There is a rapid raise in the poverty rate from 0.6 to 1.0, huge number of people are in poverty.

Female who are not working (2192) rate is more compared to self-employment (909) rate. The female having no formal saving rate is more compared to female who have formal saving. The male having no formal saving rate is more compared to male having formal saving. The poverty rate is high for people who don’t have formal savings and informal savings. People who have insurance are not under poverty and people who don’t have insurance are under high risk of poverty rate. People who have investment also face poverty and people who don’t have investment are under high risk of poverty rate. People who don’t have bank account are having high poverty probability compared to ones who have bank account.

There is a rapid raise in the phone ownership from age 17 to 33 and from age 33 to 69 the phone ownership rate has been decreased. Male use internet more compared to female. People in urban area have more internet access compared to urban area. Those who have more number of phone ownership the more the internet usage. The people who are under unemployment and other have less educational level and people who have employed and student are having higher education level 25. The dataset contains huge amount of data of age group from 18-40.

METHODOLOGY

The algorithm which is used in this study is logistic regression. The dataset contains classification records of people in poverty and people who are not in poverty hence logistic regression is used for the analysis. Logistic regression is a supervised classification algorithm and the response variable in the dataset is binomial data hence logistic regression is used for predicting the poverty. The steps which were used in implementing the logistic regression are represented in figure1

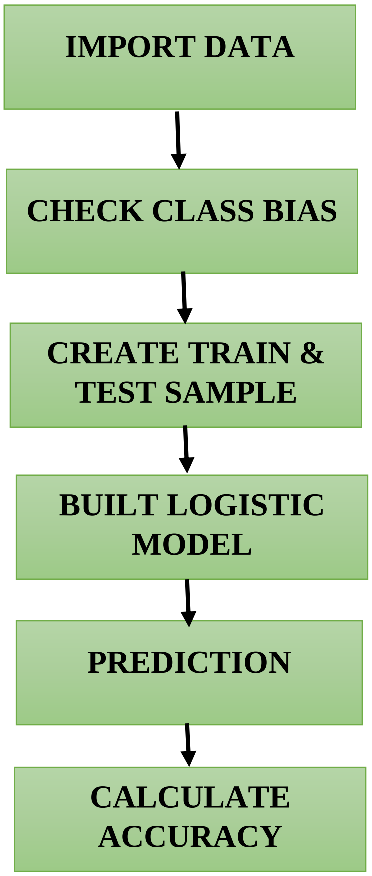

Figure 1

RESULTS

Based on the test of independence the attributes are selected for building the logistic regression model. The implementation of logistic regression gave 92% accuracy in predicting the poverty. From the analysis the attributes which are significant in identifying the people who are in poverty and not in poverty are identified. The key attributes which played major role in predicting the poverty are Literacy, Education level, informal savings, investment, bank account, urban/rural, phone, financial activities and employment.

CONCLUSION

From the analysis it is identified that people who are living in rural area are suffering under the poverty compared to the urban area. The more the educational level, the less probability of poverty rate. And also based on the employment type, the poverty probability is differing rapidly. People with bank account, investment, informal saving, high educational level and insurance are escaping from the poverty. People who are married face more poverty rate compared to the unmarried people. The only solution to overcome poverty in all countries is by means of good education irrespective of gender, which leads to respectable employment which then leads to worthy income. Well educated people will maintain formal saving from their income to overcome the poverty.

REFERENCES

[1] Narendranath, Sukhavasi, et al. "Characteristics of ‘Escaping’and ‘Falling into’Poverty in India: An Analysis of IHDS Panel Data using machine learning approach." *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018.