OR/SYST 568 – Applied Predictive Analytics

Costa Rican Household Poverty Level Prediction

Project Proposal

Introduction

The dataset[1] that has been proposed here predicts the poverty on a household basis in the country of Costa Rica. This data was made available as part of ‘data science for good’ competition on Kaggle. The objective is to use individual and household socio-economic indicators to predict poverty on a household basis. An efficient predictive algorithm which correctly classifies poverty levels would help the IDB bank (who developed the problem and provided the data) in Latin America to target poor section of the society and to identify who require the most social programs and aids. Given that an efficient algorithm is generated for this prediction, the bank claims that it can also be used to assess poverty levels across many different countries in the world in a similar fashion.

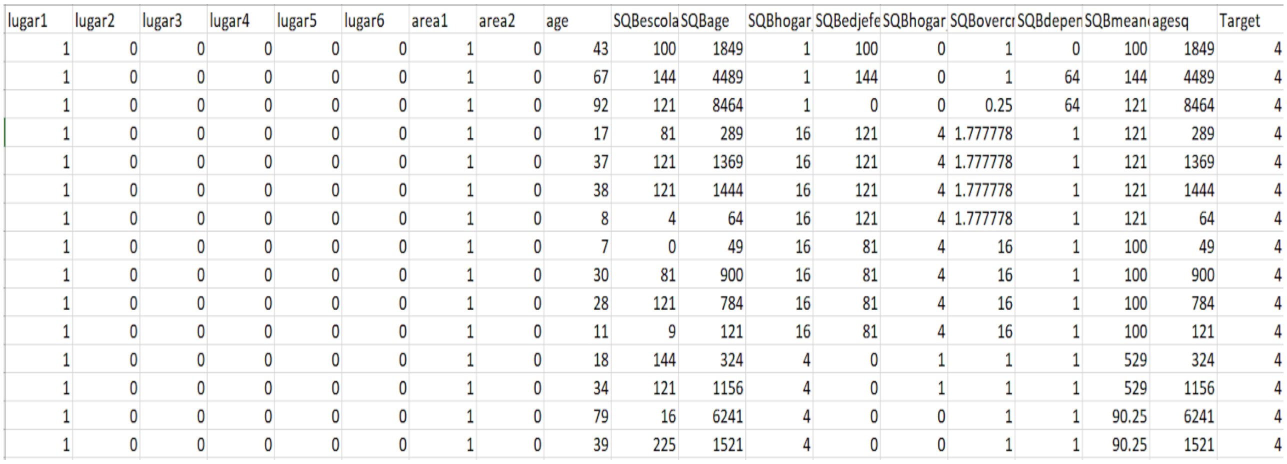
While the main objective of this dataset is to predict the level of poverty, Exploratory data analysis on the given data could answer few questions of interest such as correlation of the level of education of the individuals with other variables.

Dataset Characteristics

Based on various household characteristics, poverty in Costa Rica is categorized into four levels. The Target variables would be then classified into one of the 4 levels - Extreme poverty (1), moderate poverty (2), vulnerable households (3) and non-vulnerable households (4). The data set is split into Training dataset and Test dataset. The training set has 9557 rows and 143 columns while the testing set has 23856 rows and 142 columns. Each row in the data represents a member of the family. The training set has one additional column, Target, which represents the poverty level.

The prediction of the target variable is done at the household level. The predictor **idhogar** has the value that identifies each household uniquely. Individual’s data at each row is considered to identify derived features from the existing predictors and aggregation of individual data is to be performed to predict the target label for each household. The dataset has nominal, ordinal and ratio type columns with majority of categorical columns. Since the dataset has huge number of columns, a small subset of them along with the target variable is shown below,

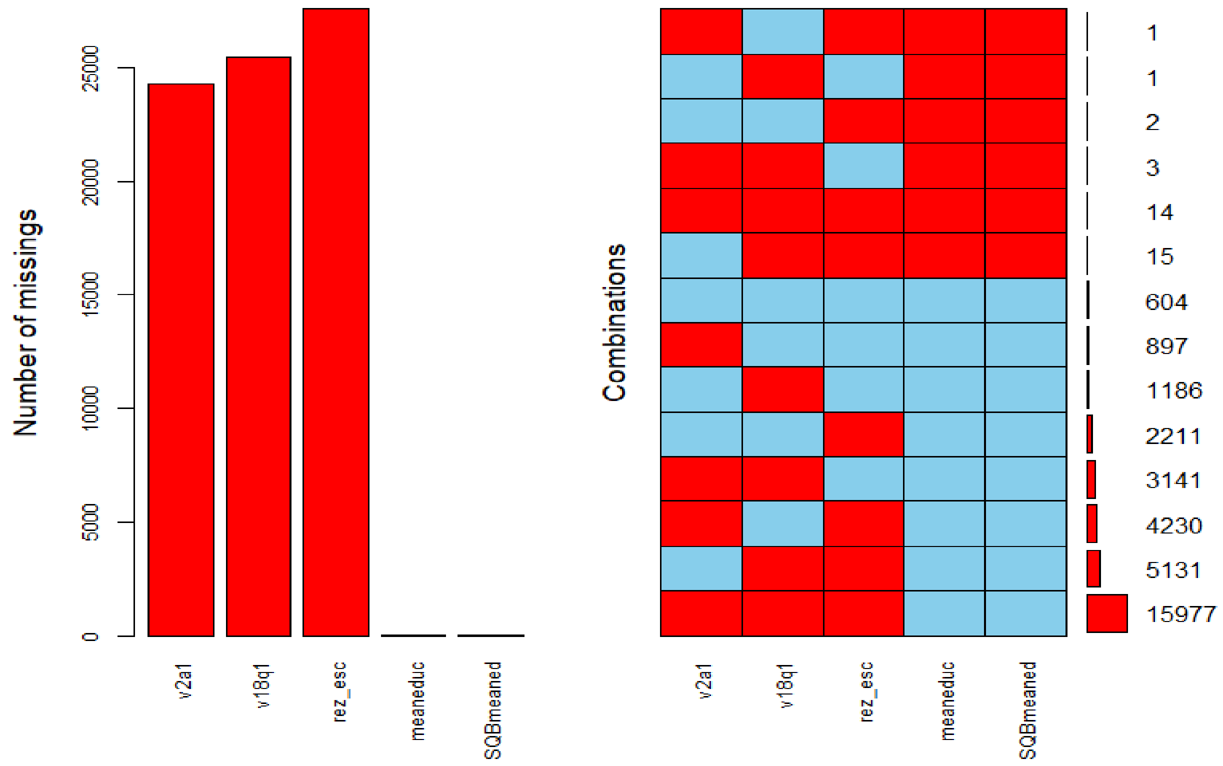
[1] "Costa Rican Household Poverty Level Prediction | Kaggle." https://www.kaggle.com/c/costa-rican-household-poverty-prediction. Accessed 7 Mar. 2019.



lugar1	lugar2	lugar3	lugar4	lugar5	lugar6	area1	area2	age	SQBescola	SQBage	SQBhogar	SQBedjefe	SQBhogar	SQBovercr	SQBdepen	SQBmean	agesq	Target
1	0	0	0	0	0	1	0	43	100	1849	1	100	0	1	0	100	1849	4
1	0	0	0	0	0	1	0	67	144	4489	1	144	0	1	64	144	4489	4
1	0	0	0	0	0	1	0	92	121	8464	1	0	0	0.25	64	121	8464	4
1	0	0	0	0	0	1	0	17	81	289	16	121	4	1.777778	1	121	289	4
1	0	0	0	0	0	1	0	37	121	1369	16	121	4	1.777778	1	121	1369	4
1	0	0	0	0	0	1	0	38	121	1444	16	121	4	1.777778	1	121	1444	4
1	0	0	0	0	0	1	0	8	4	64	16	121	4	1.777778	1	121	64	4
1	0	0	0	0	0	1	0	7	0	49	16	81	4	16	1	100	49	4
1	0	0	0	0	0	1	0	30	81	900	16	81	4	16	1	100	900	4
1	0	0	0	0	0	1	0	28	121	784	16	81	4	16	1	100	784	4
1	0	0	0	0	0	1	0	11	9	121	16	81	4	16	1	100	121	4
1	0	0	0	0	0	1	0	18	144	324	4	0	1	1	1	529	324	4
1	0	0	0	0	0	1	0	34	121	1156	4	0	1	1	1	529	1156	4
1	0	0	0	0	0	1	0	79	16	6241	4	0	0	1	1	90.25	6241	4
1	0	0	0	0	0	1	0	39	225	1521	4	0	0	1	1	90.25	1521	4

Preprocessing

The dataset requires extensive preprocessing. The missing values are visualised using aggregation plot[2].



[2] "Visualization of imputed values using the R-package VIM - R Project." https://cran.r-project.org/web/packages/VIMGUI/vignettes/VIM-Imputation.pdf. Accessed 7 Mar. 2019.

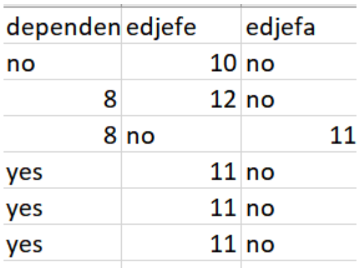
We can observe that only five columns -

1. v2a1, Monthly rent payment
2. v18q1, number of tablets household owns
3. rez_esc, Years behind in school
4. meaneduc, average years of education for adults
5. SQBmeaned, square of the mean years of education of adults (>=18) in the household

have null values. But only 3 columns (v2a1, v18q1, rez_esc) have more than 80% of the data missing. With the combinations plot, we can see that there are 14 rows in which all the five values are missing.

This makes imputation an impossible choice in them. Proper analysis has to be made if these missing values denote anything or if there is any correlation between missing values and poverty levels. Imputation is a possible choice for the other two columns with few numbers of missing values.

Apart from missing values, some columns in the dataset has a combination of numerical and binary textual values. A small subset of that kind of data is shown below.



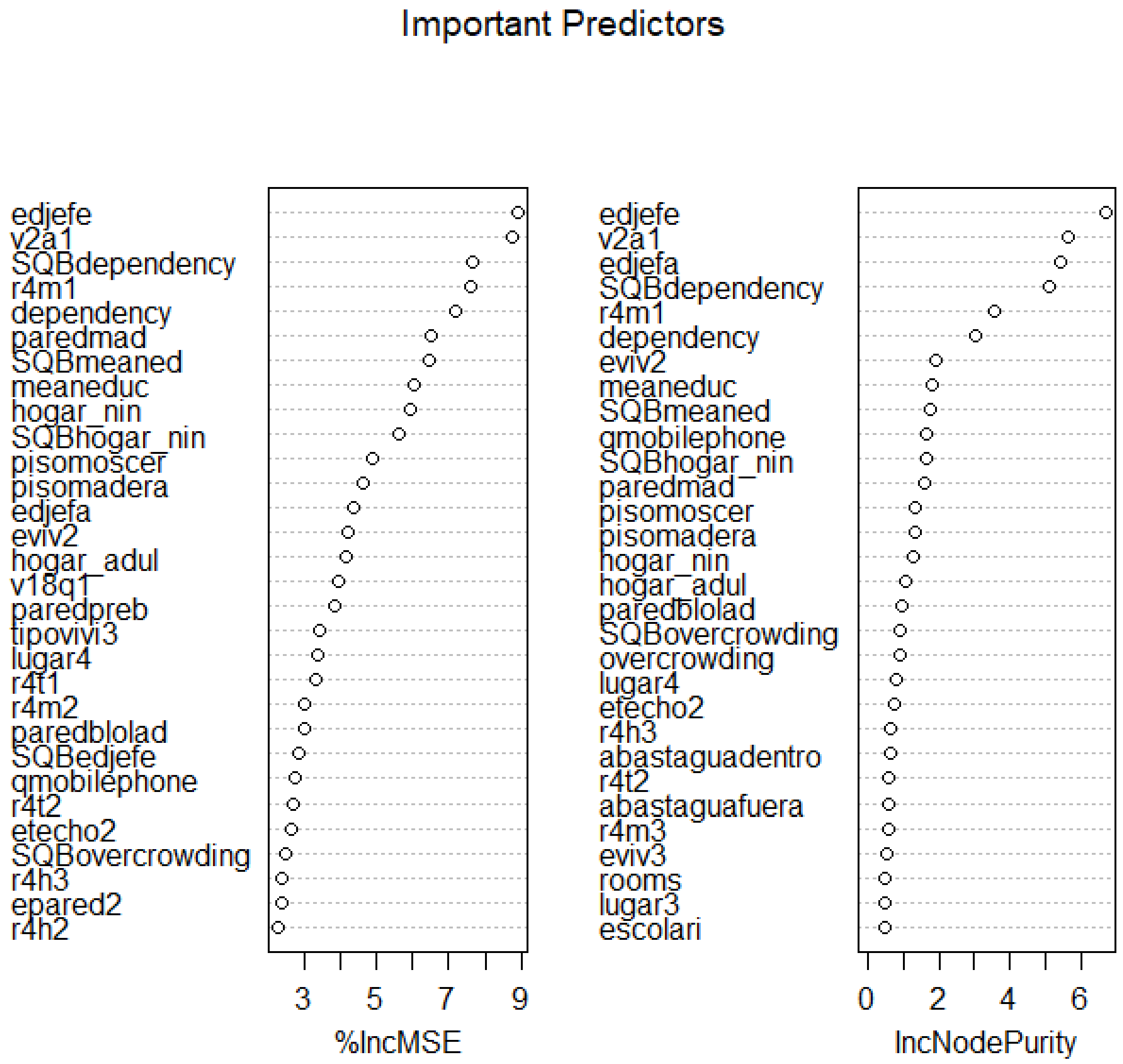
dependen	edjefe	edjefa
no	10	no
8	12	no
8	no	11
yes	11	no
yes	11	no
yes	11	no

This data could be preprocessed by attributing numerical values to those binary textual values(by factoring of levels) but the kind of numerical values that needs to be substituted could be identified only by gaining more knowledge in the domain.

Moreover, visualizing the distribution of each column could help in identifying columns with zero or near zero variance. These variables can be eliminated. Since the number of predictors is large, dimensionality reduction techniques need to be applied. Dimensionality reduction techniques along with feature engineering would reduce the predictor space.

Predictors

A random forest model would give us the outline of important features from the data. Variable importance plot without much data preprocessing is given below. From the variable importance graph given below we can observe that v2a1 which had high number of missing values is highly correlated.



Predictive models

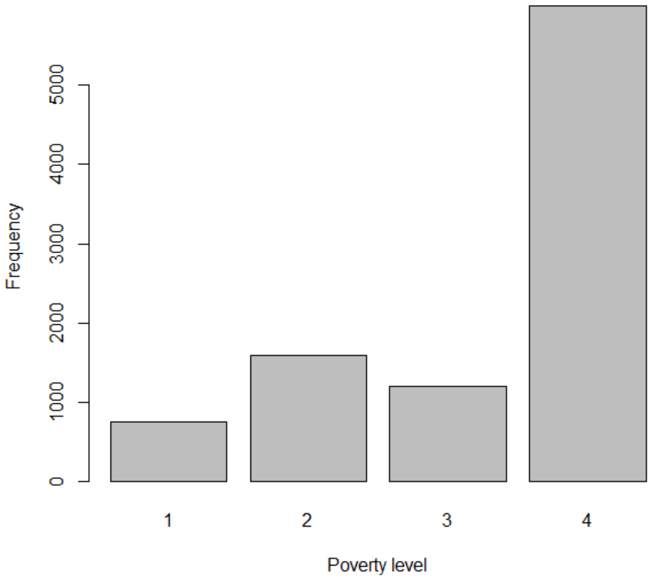
There are numerous machine learning algorithms which could be applied on this dataset. Since the target variable is categorical, linear regression models cannot be used. Some of the predictive models which are about to be employed in this data are as follows,

- Regularization models (Lasso and Ridge)
- Multinomial Logistic Regression
- Random Forest
- Support Vector Machines
- XGBoost
- LightGBM

After evaluating the performance of these models, an ensemble model comprising of other better models can also be used to improve the efficiency.

Model evaluation

A simple frequency plot on the target variable shows that more than 80% of data belongs to a certain category. The frequency plot is shown below.



Due to this unevenness of the target variable, accuracy would be a poor metric in evaluating the model. For instance, a model which selects categories would have an accuracy near 80%. Therefore, Macro F1 could be chosen as the metric to evaluate the multi classification models such as this. The macro F1 score for this dataset can be defined as[3],

$$\text{Macro F1}=(\text{F1 Class 1}+\text{F1 Class 2}+\text{F1 Class 3}+\text{F1 Class 4})/4$$

Where F1 Class1, F1 Class2, F1 Class3 and F1 Class4 are F1 scores of categories of poverty levels 1,2,3,4 respectively. This is the evaluation metric used by Kaggle to evaluate models for this competition.

Other possible metric for the model evaluation is the micro average. While macro-average computes the F1 score independently for each class and then takes the average, micro-average aggregates the contributions of all classes to compute the average metric. In a multi-class classification model if the distribution of target variables are significantly skewed, micro-average metric is also a preferred choice.

[3] "A Complete Introduction and Walkthrough | Kaggle." https://www.kaggle.com/willkoehrsen/a-complete-introduction-and-walkthrough. Accessed 7 Mar. 2019.